INEQUALITY UNEXPLAINED: AN EXPERIMENTAL TEST OF DIFFERENTIAL TREATMENT IN SCHOOL DISCIPLINE DECISION-MAKING

Sarah McGill Davis

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for the degree of
Master of Arts in the Department of Sociology.

Chapel Hill
2018

Approved by:

Karolyn Tyson

Katherine Weisshaar

Ted Mouw

ABSTRACT

Sarah McGill Davis: Inequality Unexplained: An Experimental Test of Differential Treatment in School Discipline Decision-Making
(Under the Direction of Karolyn Tyson and Katherine Weisshaar)

Research consistently shows a high level of racial disparity in public school discipline whereby white students receive fewer and less severe disciplinary actions than their black classmates, even for the same offense (Losen and Gillespie 2012; Fabelo et al. 2011; Fenning and Rose 2007). This study employs a survey experiment to test whether this disparity is driven by differential treatment. U.S. public school educators were presented with hypothetical student misbehavior vignettes and asked their disciplinary recommendation. I control for student behavior while varying student race and gender in order to determine whether educators recommend differential discipline across student race and gender categories. Multivariate analyses of the survey experiment responses fail to provide evidence for the differential treatment hypothesis. I discuss these findings in the context of the current literature, offering four possible explanations for the study's null results.

ACKNOWLEDGEMENTS

I would like to extend my sincere thanks to the joint chairs of my Master's committee, Karolyn Tyson and Katherine Weisshaar, for their thorough, insightful guidance throughout my thesis-writing process. I am also grateful to Ted Mouw who served on my Master's committee and provided helpful feedback on the project. My friends and colleagues in Dr. Tyson's advising group and those in the Race Workshop offered me valuable input and much-needed comradery during the years spent on this project. Finally, I am indebted to my fiancé Nathan Brown for his unwavering support and encouragement through the ups and downs of graduate school.

TABLE OF CONTENTS

INTRODUCTION

Research consistently shows a high level of racial disparity in public school discipline whereby white students receive fewer and less severe disciplinary actions than their black classmates, even for the same offense (Losen and Gillespie 2012; Fabelo et al. 2011; Fenning and Rose 2007). This disparity occurs not just in the realm of office referrals for minor incidents but also out-of-school suspensions and expulsions, jointly termed "exclusionary discipline" (Fenning and Rose 2007). In US public schools, white students are suspended or expelled one-third as often their black counterparts; in the 2011-2012 school year, only 5% of white students across the country faced suspension while 16% of black students experienced this same punishment (US Department of Education 2014). Since early studies conducted in the 1970s, these disparate rates of punishment between white and black students have only grown (Fabelo et al. 2011; Rocque 2010; Kinsler 2011). School discipline shows yet more disparity at the intersection of race and gender, whereby boys and girls within each racial category experience punishment at differential rates and for differential reasons. While boys overall face higher rates of discipline than girls, the racial discipline gap among girls is substantially larger than the racial discipline gap for boys, suggesting that discipline disparities are simultaneously shaped by gendered and racialized ideals of student behavior (Blake et al. 2010; Wallace et al. 2008; Morris and Perry, 2017; Skiba et al. 2002). Black girls, compared to other race and gender groups, are significantly more likely to be disciplined for disruptive behavior, dress code violations, disobedience, and aggressive behavior, presumably related to their perceived failure to follow traditional views of femininity (Blake et al 2010; Morris 2005)

Exclusionary discipline has wide-ranging and detrimental ramifications for students (Fenning and Rose 2007; Fabelo et al. 2011; Losen and Gillespie, 2012; Skiba et al. 2014). During the 2009-2010 school year, for example, over three million children in grades K-12 lost instructional time due to out-of-school suspension (Losen and Gillespie, 2012), which often leaves students unoccupied and unsupervised at home. The compounding impacts of exclusionary discipline can lead students to become academically disengaged and experience lower educational achievement over time (Skiba et al. 2014). Students who receive exclusionary discipline are more likely to be held back a grade or drop out of high school (Fabelo et al. 2011). Exclusionary discipline can also result in decreased self-esteem and negative self-image, as well as other negative emotional and psychological impacts (Ferguson 2000). These factors may then contribute to involvement with the juvenile justice system and incarceration in adulthood, a process many refer to as the "school-to-prison pipeline" (Fabelo et al. 2011; Fenning and Rose, 2007; Skiba et al. 2014).

It is important to note that it is not discipline in general that researchers find problematic, but specifically exclusionary discipline that removes students from the school environment (Skiba et al. 2008). Exclusionary discipline shows negative outcomes for all students, but the problem takes on added weight when some populations face a higher risk than others, especially when one acknowledges that the students subjected to harsher discipline are often those already facing difficulties in other aspects of their lives. Thus, rather than addressing the needs these students present, schools in fact exacerbate existing inequalities and magnify the likelihood of later-in-life difficulties for these students (Rocque 2010; Ferguson 2000).

Scholars continue to debate what drives disparities in exclusionary discipline and how these disparities may be diminished. As will be discussed in the literature review, some scholars

suggest that differential discipline represents an otherwise proportional response to differential rates of misbehavior, either due to black and white students' purportedly differential inclinations toward misbehavior (Rocque 2010; Kinsler 2011; Kinsler 2013) or operating indirectly through black and white students' differential socioeconomic or class differences (Skiba et al. 2002; Fabelo et al. 2011; Rocque and Paternoster, 2011). Previous research shows mixed evidence for these hypotheses; meanwhile, other scholars' work lends credence to the alternate hypothesis that differential discipline results from differential treatment of black and white students (Skiba et al. 2002; Rocque 2010; Wallace et al. 2008; Morris 2005; Morris and Perry, 2017; Fabelo et al. 2011).

To contribute to this debate, this study uses experimental methodology to test the differential treatment hypothesis. I control for student behavior to determine whether educators systematically recommend more lenient disciplinary actions for white students. Additionally, I test for relationships between educators' disciplinary recommendations and two related factors: 1) the degree of discretion permitted to educators in making their disciplinary recommendations and 2) the rationales that shape educators' reasoning. Multivariate analysis is conducted to assess the degree to which race and gender shape participants' disciplinary responses. Findings from this study fail to provide evidence for the differential treatment hypothesis, suggesting that instead, participants recommend nearly identical disciplinary responses across race and gender categories. I discuss these findings in the context of the current literature on the differential behavior and differential treatment hypotheses and conclude the paper with four possible explanations for the study's null results. This paper opens up a host of questions, both methodological and substantive, for further exploration.

Literature Review

This study is situated within an existing body of empirical and theoretical scholarship that addresses both the scope and nature of school discipline disparities. I begin this section with a review of that scholarship, followed by a review of the existing research on the ways in which racialized discipline disparities are gendered. Finally, I summarize existing hypothesized explanations for these disparities, several of which this study will test empirically.

Existing literature demonstrates persistent patterns of exclusionary discipline that appear to privilege white students while black students face grave disadvantages (Losen and Gillespie 2012; Fabelo et al. 2011; Fenning and Rose 2007). In 2009-2010, for example, one in six black students across the country experienced exclusionary discipline while only one in twenty white students faced this risk (Losen and Gillespie 2012). The use of exclusionary discipline and the degree of racial disparity in school discipline have both increased since the 1970s (Skiba et al 2011; Losen and Gillespie 2012; Kinsler 2011), and the disparity extends beyond suspensions and expulsions to include disciplinary office referrals (Skiba et al 2011; Rocque 2010). For example, Skiba et al (2011) conducted descriptive and logistic regression analyses of elementary and middle school office referral data and found that African American children are 2.19 (elementary) to 3.78 (middle) times more likely to be referred to the office for what is perceived as problematic behavior.

School discipline disparities between white and black students have not only existed but widened in the years following school desegregation. Kinsler (2011) finds that suspension rates for white students rose by 64% from 1972 to 2000, but 120% for black students. Scholars argue this increase stemmed from the "tough on crime" policies of the 1970's and 80's, the increased media representations of black youth as criminals, and the Gun Free Schools Act of 1994 which

led to the popularization of so-called zero-tolerance policies (Fabelo et al. 2011). Fabelo and colleagues (2011) explain that, while zero-tolerance policies were initially designed to incentivize the swift expulsion of students who brought guns to schools, many school districts acted on their own volition to expand the scope of these policies beyond the mandatory standards. As of the 1990's, zero tolerance policies were widely adopted across U.S. public schools, mandating swift and severe punishments without consideration of the extenuating circumstances, contextual factors, or the gravity of the behavior itself (Skiba et al. 2008:852). Fabelo et al. (2011) report that, as of 1997, 79% of public schools across the country utilized zero-tolerance policies for not just weapons but alcohol, drugs, and violence. This has led to a steep increase in suspensions, from 1.7 million suspensions in 1974 to 3.3 million suspensions in 2006, representing an increase from 3.7 to 6.8 percent of all U.S. public school students. This new wave of suspensions includes noteworthy stories that caught media attention: a ten-year-old girl, for example, was expelled after she brought a small knife in her lunchbox, packed by her mother so she could cut an apple; another story involves a young teenager who was expelled for breaking school rules by talking on the phone to his mother, a soldier deployed in Iraq (Skiba et al. 2008). Sadly, the bulk of these increased suspensions and expulsions were assigned to black students.

While some policymakers proposed that zero-tolerance policies would lead to consistency in the distribution of punishments, research shows that racial disparities not only persisted but increased with their implementation (Skiba et al. 2008:854; Fabelo et al. 2011). Thus, rather than improving consistency in the application of school discipline, zero tolerance only served to increase the severity of the sanctions rather than evening their distribution (Skiba et al. 2008; Fabelo et al. 2011). The most recent data on school discipline disparity shows

widespread evidence for an increase in the magnitude of racial disparities in school suspensions and expulsions that corresponds with zero-tolerance, "tough on crime" policies and practices.

Losen and Gillespie's (2012) descriptive analysis of suspensions revealed that use of suspension varies drastically by state and by district, as does the degree of racial disparity. North Dakota suspends students the least of all the states, with 2.2% of all students experiencing suspension in the academic year, while South Carolina had the highest suspension rate of 12.7%. Of the 6,779 districts included in their analysis, 839 districts suspended 10% or more of their student body at some point during the school year, and nearly 200 districts suspended more than 20%. Racial disparity in suspension rates is also notably varied across districts. More than 300 districts suspended 25% or more of their black students; meanwhile over 1,400 districts suspended less than 3% of their black students. The highest-suspending district for black students is Pontiac City, Michigan that suspended 67.5% of their black student body. Despite this variation across states and districts, a higher proportion of black students are suspended in every state, suggesting persistent patterns of racial disparities across the country.

Relatively few studies have considered school discipline at the intersection of race and gender, yet scholars continually highlight the need for more nuanced understandings of the ways that race and gender interact as overlapping oppressions (Morris 2016; Collins 2009; Ferguson 2000). In her 2016 book *Pushout: The Criminalization of Black Girls in Schools,* Monique Morris raises concern about the limited attention to black girls' experience in schools, experiences that are often shaped in troubling ways by surveillance, ostracism, and marginalization. Although black male students have received comparatively more attention with regards to school discipline, Morris claims that the omission of black girls from the existing research "obfuscates the ways in which Black females and males experience this phenomenon

together *and* differently" (Morris 2016:9), thereby hindering our sociological understanding of the complex workings of school discipline disparities. Patricia Hill Collins (2009) also underscores the need for female Black voices and experiences to enter social science scholarship. She says their omission from the literature is "...neither accidental nor benign", claiming furthermore that the continual invisibility of Black women and girls has in fact served to maintain social inequalities (Collins 2009:5). Not only is this omission unjust, but it is myopic. Collins explains that the experiences of oppressed people have a unique vantage point from which to understand their oppression (Collins 2009:11); thus, the inclusion of African American females in studies of school discipline disparity is critical for a clearer understanding of intersecting oppressions of race, class and gender in the American public school system.

Jamilia Blake and her colleagues (2010) were among the first to conduct research on Black girls' experiences of school discipline independent of the experiences of Black boys. Through their analysis of a Midwestern school district's discipline records from 2005-2006, they found that the racial discipline gap is not only present for girls, but it is in fact wider than the racial gap previously reported for boys (Skiba et al. 2002; Blake et al. 2010). In a related study, Edward Morris and Brea Perry (2017) analyzed detailed, longitudinal school discipline records of over 30,000 middle and high school students in a Kentucky district from 2007-2011. After controlling for socioeconomic status, school location, special education status, and academic achievement, their analysis shows that Black girls are three times as likely to receive an office referral as white girls, representing a racial disparity even greater than that between black and white boys. Wallace and colleagues (2008) conducted an analysis of tenth graders' self-reported behavior and discipline experiences and found that, controlling for family structure, parental education, and the region and urbanicity of the school, black boys are 3.3 times more likely than

white boys to be suspended or expelled, compared to black girls who are 5.4 times more likely to be suspended or expelled than white girls.

Scholars have also documented differences in the types of behaviors girls and boys are typically disciplined for, compared across gender and across race and ethnicity. Blake and colleagues (2010) found that black girls, compared to white and Hispanic girls, were more likely to be referred for 1) defiance, 2) inappropriate dress, 3) using profane language toward a student, and 4) physical aggression. Similarly, Morris and Perry (2017) found that black girls are significantly more likely to receive office referrals for disruptive behavior, dress code violations, disobedience, and aggressive behavior. Scholars suggest that Black girls' higher rates of discipline are related to their perceived failure to follow traditional views of femininity (Blake et al 2010; Morris 2005). In sum, empirical research clearly shows that race-based discipline disparity is shaped by dynamics of gender. The next section will address what is currently known and unknown with regards to the factors driving these differential rates of discipline across race and gender.

Scholars continue to debate whether differential punishment is driven by "legitimate" or "illegitimate" factors (Rocque 2010:573). "Illegitimate" factors would include bias and discrimination, whether conscious or unconscious, intended or unintended, that lead to the differential treatment of students. "Legitimate" factors would include the possibility that certain groups of children do in fact misbehave more than other groups of children, and thus their higher rates of discipline are warranted. Amidst this debate, the possibility of differential behavior represents an important factor to consider when assessing the reasons behind discipline disparities.

Interestingly, scholars have found evidence for differential behavior between girls and boys (Skiba et al 2002), suggesting the discipline gap between boys and girls is at least partly driven by "legitimate" reasons. Russell Skiba and colleagues (2002) collected discipline records from a large, urban Midwestern district and conducted discriminant analysis of the behaviors for which students received disciplinary office referrals. They found that boys significantly more likely to be referred for eleven out of the twelve possible categories of behavior infractions, including throwing objects, spitting, fighting, vandalism, and sexual acts. In contrast, girls were more likely to be reported for only one behavior violation: truancy (2002:331). It is important to note that, although the gendered differences in reported behavior were statistically significant, this only allows us to conclude that boys and girls are *reported* for significantly different behaviors. Additional evidence is needed to conclude with certainty that boys' and girls' behaviors are truly different. Furthermore, while these data provide some evidence for differential behavior by gender, the findings are complicated by Wallace et al.'s data (2008) showing that black boys' rates of suspension and expulsion are 3.3 times that of white boys, while for black girls their suspension and expulsion rates are five times that of white girls. Clearly, gendered differences in misbehavior and discipline are not operating identically across racial categories. Next, a review of the literature on differential behavior by race is provided, showing that while there is some limited evidence for the existence of differential behavior along gendered lines, studies assessing differential behavior by race are much less conclusive.

One hypothesis for racial discipline disparities is that differential rates of misbehavior are actually operating through differential rates of poverty experienced by black and white students, making race spurious. In other words, considering that children who live in poverty are found to be more likely to misbehave, and considering that race and poverty are closely linked in US

society, some hypothesize that this explains why black children receive more discipline (Skiba et al. 2002). However, multivariate analysis of student demographic data and discipline records shows that these poverty-based explanations do not account for the degree of racial disparity exhibited in schools (Skiba et al 2002; Vavrus and Cole 2002; Welch and Payne 2010; Fabelo et al. 2011). Skiba and colleagues (2002) conducted multivariate logistic regression of school discipline records and found that race is in fact a stronger predictor of suspensions and expulsions than socioeconomic status, thereby effectify nullifying the hypothesis that socioeconomic disparities can explain racial disparities in school discipline. Similarly, Fabelo et al (2011) compared exclusionary discipline rates for white and black middle schoolers in Texas, controlling for socioeconomic status as well as school attendance and academic performance, and found that black students with otherwise similar attributes were still 31% more likely to be disciplined than students of other races.

Alongside the poverty-based hypothesis is the hypothesis that black students inherently misbehave more, resulting in their higher rates of punishment (Skiba et al. 2002). The lack of observational data on school discipline disparities makes it difficult to test directly for differential behavior. Several studies, however, seek to test for differential behavior by analyzing school records of the student behaviors that elicited disciplinary responses, racial differences in teachers' assessments of their students' behavior, or students' own reports of their behavior. Findings from each of these three methodological approaches are reviewed below, demonstrating that although there is some limited evidence for differential behavior by race, it is not sufficient to account for the size of the discipline gap (Skiba et al. 2002; Skiba et al. 2011; Skiba et al. 2014; Roque 2010).

In one methodological approach, scholars have attempted to test for the differential behavior hypothesis by statistically comparing black and white students' rates of reported misbehavior across several categories, from low to high severity (Skiba et al. 2002; Skiba et al. 2011; Skiba et al. 2014). In one key study, Skiba et al (2002) collected discipline data from 19 middle schools in a large, Midwestern school district. They tested for differential behavior indirectly by analyzing the types of behavior for which students received disciplinary office referrals. They found that, although black students received a significantly higher proportion of disciplinary referrals than other racial groups, there was no evidence that they were referred for more serious forms of misbehavior. Instead, they found that black students were significantly more likely to be referred for behaviors that were subjectively interpreted, including "disrespect, excessive noise, threat, and loitering" (2002:332) compared to the more objectively interpreted behaviors that white students were referred for: "smoking, leaving without permission, vandalism, and obscene language" (2002:332). These data suggest that black students' higher rates of office referrals are not due to differential behavior, but are instead due to educators' differing propensity to refer black students, a form of differential treatment that will be discussed in the next section.

In another key study, Russell Skiba collaborated with a different group of colleagues in 2011 to analyze discipline records from a national sample of 436 elementary and middle schools, including data on office referrals and other disciplinary sanctions. From these data, Skiba and his colleagues measured correlations between the type of infraction that provoked the office referral and the disciplinary response. Descriptive analysis and logistic regression showed that black students were significantly more likely than white students to receive office referrals, and once they were referred to the office, they were more likely to receive suspensions and expulsions for

the same behaviors exhibited by other racial groups. Similar to Skiba and colleague's previous work (2002), this study found that black students were significantly more likely to be referred for disruption and noncompliance, behaviors that are subjective and interactive in nature.

Michael Rocque (2010) took a different methodological approach to assess the differential behavior hypothesis, relying instead on classroom teachers' assessments of their students' behavior. He gathered disciplinary and behavior records from 45 elementary schools in Virginia, accounting for nearly 29,000 students. To measure behavior, Rocque asked teachers to report for each of their students "how much the student acts out, disregards rules, and is generally disruptive" (567), generating a measure of what is termed "externalizing or antisocial" behavior (2010:567). Rocque then used teacher-reported behavior as a control in his analysis to test whether race predicts disciplinary office referrals. Rocque also controlled for a number of other factors including socioeconomic status, gender, age, special education, and academic performance. He found that when then teacher-assessed behavior is not included in the model but all other controls are, black students have a 2.27 greater odds than other racial groups of receiving an office referral. However, when teacher-assessed behavior is added to the model, black students have a 1.58 greater odds of receiving an office referral compared to other racial/ethnic groups. Rocque concludes that differential behavior does account for some degree of the racial disparity in office referrals. However, it is important to note, as Rocque does, that accounting for behavior in this way is not a fully objective measure since teachers may hypothetically exhibit bias in their assessments. The final 1.58 odds ratio suggests that, even when teacher-assessed behavior is included in the model, black students still receive significantly more office referrals. This lends further evidence for the existence of bias or differential treatment at the hands of school administrators, which the proposed study will empirically test.

Wallace et al (2008) added a new layer to the question of differential behavior by collecting students' self-reports of their own behavior and discipline incidents. As summarized above, they found significant variability in the rates at which black and white boys and girls report receiving discipline, suggesting that discipline disparities are not only functioning along the lines of race but also by gender. To test for differential behavior, Wallace and colleagues were particularly interested to compare the rates at which each of these demographic groups reported engaging in serious zero tolerance-worthy offenses. Interestingly, they found small but significant differences in students' self-reports of their own zero-tolerance-worthy behavior, including the proportion of students who report having brought a gun to school. Their data shows that black students are significantly more likely than white students to carry a gun to school: 7.9% of black boys compared to 3.0 % of white boys responded that they had carried a gun to school at least once in the past four weeks, alongside 1.7% of black girls and 0.5% of white girls who reported this same behavior. While Wallace and colleagues' data is perhaps the closest to an objective measure of differential behavior, the sheer numbers they report are unsettling; if taken at face value, these proportions suggest 382 black boys, 702 white boys, 91 black girls, and 122 white girls had brought a gun to school in the past four weeks out of their total sample of 74,000 tenth graders. However large these numbers may seem, Wallace et al report that, although the difference in students' self-reports of this behavior are statistically significant, the apparent behavior disparities are in effect "relatively small" (2008:53), considering their finding that black boys are 3.3 times more likely than white boys to be suspended or expelled and black girls are 5 times more likely than white girls to be suspended or expelled. Thus, Wallace et al conclude that behavior disparities are "insufficient to account for the relatively large racial and ethnic differences in school discipline" (Wallace et al 2008:10).

In all, a review of the literature reveals that while there is some limited evidence for differential rates of misbehavior between black and white students, these differences are not large enough in magnitude to account for the full scope of the disparity (Skiba et al. 2002; Wallace et al 2008; Rocque 2010; Skiba et al 2011). Scholars infer that, rather than either differential behavior or differential socioeconomic status driving racial disparities, it is differential treatment by teachers and administrators that primarily contributes to these disparate outcomes (Losen and Gillespie 2012; Skiba et al 2014; Fabelo et al 2011; Ferguson 2000; Welch and Payne 2010; Lewis and Diamond 2015; Vavrus and Cole 2002). In the next section, I review evidence for the role of differential treatment and bias in school discipline decision-making, what Rocque (2010) would term the "illegitimate factors" contributing to the racialized and gendered discipline gap, and the questions that remain in the literature.

Alongside the mixed evidence for differential behavior explanations is a more robust and consistent set of findings from qualitative studies suggesting differential treatment; however, questions remain as to the nature of the differential treatment and how it may be reduced. Discipline disparities, some scholars argue, stem from educators' desire for control in the classroom and the accompanying fear of losing control, desires that may be unevenly directed toward the behavior of black students (Fenning and Rose 2007). In contrast to white students whose cultural and linguistic expression is highly regarded (Ferguson 2000), poor students of color are often viewed as "not fitting in" with school norms and perceived as "dangerous" or "troublemakers" (Skiba et al 2002).

Furthermore, differential treatment that is racial in nature is also gendered. Monique Morris's ethnographic work (2016) illuminates the ways that identity politics shape perceptions and expectations of girls and boys in school settings. Morris found that black girls are often

categorized dichotomously as either “good girls” or “ghetto girls” (Morris 2016:10). She contends that students perceived as “good girls” display behaviors that align with a “narrow, White, middle-class definition of femininity” while “ghetto girls” are instead “frequently labeled as nonconforming and thereby subjected to criminalizing responses” (Morris 2016:10). Thus, her ethnographic work documents educators’ use of school discipline to regulate Black girls’ behavior and identities. Ferguson’s ethnographic study (2000) revealed the ways discipline dynamics are shaped by racialized perceptions and expectations of masculinity. She found that black boys’ behavior infractions are readily interpreted as ominous and willful, in contrast to the childlike innocence and naiveté that is attributed to their white peers. These cognitive frames, alongside the desire of school staff to remain in control of student behavior, result in black male students being removed for infractions at higher rates.

Edward Morris’s ethnographic study in a Texas middle school (2005) came to a similar conclusion, finding that race took on different meanings for boys and girls. For black boys, blackness was seen as dangerous and threatening while for girls, their blackness was seen as insufficiently feminine. Thus, educators restricted black girls’ behavior that was seen as loud and aggressive and sought to shape them into “ladylike” beings by enforcing quiet, compliant, and deferential behavior. “Brownness” or Latino/a identity was also perceived differently along gendered and class lines. Latina girls were not perceived as insufficiently feminine; however, Latino boys were seen as potentially aggressive and threatening, reflecting popular culture that portrays Latino masculinity as dangerous. These differences in perceptions led to Latino boys being closely monitored and disciplined similar to black males. However, unlike black males, Latino boys could overcome this through displays of cultural capital through forms of dress and manners. Meanwhile, whiteness and Asianness seemed to operate as protective labels. White and

Asian students were not held to as close scrutiny or discipline as black and Latino students. Furthermore, unlike Latino boys, white and Asian students were not required to use dress and manners to overcome surveillance and perceived threat. The author explains, "Even when [white and Asian students] affected a street style almost identical to that of black and Latino youth, educators typically interpreted white and Asian American boys as harmless and white girls as well mannered" (Morris 2005:45). They found that Whiteness and Asianness seemed to signal "docility and normative masculinity and femininity" (Morris 2005:45).

If in fact differential treatment is what primarily drives discipline disparity, the degree of discretion permitted to decision-makers when making their discipline decisions could potentially widen or narrow the gap. Fabelo et al.'s longitudinal analysis of public school students in Texas (2011) compared discretionary discipline referrals to referrals mandated by the state policy and found that just three percent of suspensions and expulsions were for behaviors that state law mandates these punishments. The remaining 97% of suspensions and expulsions were made at the discretion of school staff for violations of school conduct codes (2011:x). These scholars found that white, Hispanic, and African American students faced mandated discipline at comparable rates, but their rates of discretionary discipline were significantly different. Similarly, Skiba et al (2002) identified a pattern whereby black students were disciplined more often for subjective or interpretive behavior infractions. Together, this research suggests that when minor behavior infractions such as classroom disruptions or non-compliance are used as discretionary grounds for suspensions, this creates room for possible bias and differential treatment.

The current scholarly understanding of the role of discretion in school discipline disparity is limited. However, research shows discretion plays a part in racially disparate

recommendations for exceptionality testing for students (Fish 2016), as well as many other spheres beyond the realm of education, including foster care placements (Doyle 2008), custody disputes (Perry 1991), and hiring decisions (Gaddis 2015; Correll 2013). Correll (2013) finds that the degree of discretion permitted in hiring decisions in large corporations makes room for bias, whereas clear, formal policies minimize room for bias and thereby lead to reduced inequality. To decrease the resulting inequity, Correll urges institutions to narrow the degree of discretion permitted, thereby reducing the room for stereotypes and bias to influence decision-making. Lewis and Diamond (2015) offer a similar suggestion, recommending tightening up policies to narrow the gap between the ostensive and performative aspects of disciplinary routines.

Rather than enforcing fair, equitable treatment, organizational mechanisms exist within schools that perpetuate inequity under a veil of race-neutral policies. Lewis and Diamond (2015) explain that this disconnect happens in the gap between the "ostensive aspect" and the "performative aspect" of school discipline, the ostensive being the apparently race-neutral school discipline policy, and the performative being the way the policy is implemented in practice. Through ethnographic and interview research, Lewis and Diamond identify a range of ways in which ostensive and performative routines differ, bringing about differential selection and differential processing of disciplinary incidences for black and white students. Both differential selection and differential processing, they assert, may be exacerbated by the degree of discretion permitted by the school discipline policy.

These studies suggest pivotal possibilities to minimize education inequalities by reducing the degree of discretion permitted in discipline decision making; however, current scholarship has not yet tested this relationship. This study represents the first experimental assessment of

differential treatment in school discipline decision-making, a critical addition to existing research because it allows for the isolated manipulation of race, gender, and discretion in otherwise identical student behavior scenarios.

Questions and Hypotheses

The current study uses experimental research methodology to assess the potential role of differential treatment in generating disparities in school discipline and the degree to which differential treatment may be shaped by a variety of contextual factors. The following research questions and hypotheses will be tested:

1) How does the severity of disciplinary actions change depending on students' race? Is there evidence for differential treatment by student race? If so, how is differential treatment gendered?

I hypothesize that participants will prescribe harsher discipline to black students than to white students, and that these differences will vary at the intersections of race and gender. This hypothesis is consistent with empirical research that demonstrates persistent patterns of disparity whereby white students receive fewer and less severe punishments than their black peers, and that rates of discipline are typically highest for black boys, followed by black girls, white boys, and white girls in sequence (Wallace et al, 2008). If this hypothesis is supported by the data, it will provide evidence for differential treatment as a causal factor for the discipline disparity, and furthermore, will demonstrate its raced and gendered nature.

2) How do racial variations in the severity of disciplinary action change depending on the degree of discretion permitted by school policies?

I hypothesize that when participants are given a higher degree of discretion, their disciplinary decisions will show higher rates of disparity across the disciplinary actions they recommend to black and white boys and girls. This hypothesis is derived from research that shows a wide gap between the "ostensive" and "performative" aspects of school discipline (Lewis and Diamond 2015) and that document higher degrees of disparity when decision makers have more latitude within their disciplinary purview (Fabelo et al, 2011; Skiba et al, 2014).

3) What rationales shape educators' decisions and how do these shape the severity of their disciplinary recommendations for black and white students?

I hypothesize that respondents will tend to punish white students, and especially white girls, with rationales that involve preventing future misbehavior and restoring good behavior. For black girls, rationales will show particular evidence of respondents' desires to instill "properly feminine" behavior including being quiet, still, and unobtrusive. In contrast, respondents will tend to punish black students, and black boys in particular, with rationales that involve protecting other students and staff and minimizing disruptiveness. This hypothesis is based on empirical research showing racialized and gendered interpretations of and reactions to student behavior that tend to classify black boys as threatening and dangerous, black girls as overly sexual and aggressive, and white girls as naive and in need of protection (Morris, 2005; Morris, 2016; Ferguson, 2000).

RESEARCH METHODS

This study tests for differential treatment of white and black girls and boys in discipline decision-making and seeks to identify which if any conditions and rationales shape educators' disciplinary recommendations. To do so, this study employs an experimental survey design, a methodology that allows researchers to manipulate the variables of interest while isolating and controlling a range of other factors.

Experimental Design

While observational studies can accurately determine the factors that correlate with various forms of social inequality, they are limited in their ability to directly measure the potentially causal role of discrimination in generating these forms of inequality (Gaddis 2015). In contrast, experimental studies allow researchers to identify possible causal factors through the isolated manipulation of specific factors (Levin 2011). In recent years, experimental methodology has been used in the form of audit studies and survey experiments to make important strides in understanding the mechanisms that drive disparities, including those in the labor market and in public K-12 education (Pager 2003, Gaddis 2015, Weisshaar 2018, and Fish 2016). A brief review of this literature is provided next, illustrating the ways experimental methodology has been critical in pushing forward a sociological understanding of the mechanisms driving social inequality.

Devah Pager's audit study "The Mark of a Criminal Record" (2003) represents a critical contribution to the current academic understanding of racial inequality in the labor market. In

this study, Pager recruited matched pairs of men to apply in person to 350 entry-level jobs across Milwaukee, varying the applicants' race (black or white) and whether they indicated that they had formerly been incarcerated. Analysis of the employers' responses revealed that, while having a criminal background resulted in fewer call-backs for both black and white applicants, its effect was 40% larger for black applicants than for white applicants. In fact, formerly incarcerated applicants who were white were more likely to receive a call back than black men who had never been incarcerated. Because the applicants were different only in their race and their criminal background, this study provides strong evidence for racial discrimination in hiring practices.

In another key labor market study, Michael Gaddis (2015) conducted an experimental study in which he created fictitious résumés, varying only the race of applicant and the college from which they graduated, and submitted matched pairs of résumés to 1,008 online job postings. Results show that employers were significantly less likely to respond to black candidates than to white candidates, and when employers did respond to black candidates, they were for lower-paid and less prestigious jobs. Even when black candidates had degrees from elite universities, their likelihood of receiving a response was only as high as those of white candidates from less selective universities. Because the only difference among the candidates was their race and the university from which they graduated, this study provides strong evidence for discrimination in hiring practices, showing that even a degree from an elite university does not protect black job candidates from racial discrimination.

Other scholars have used experimental methodology to test for gender discrimination in the labor market. Katherine Weisshaar (2018) used a survey experiment to compare the ways that mothers and fathers were perceived when re-entering the workforce after a lapse in employment, either due to job loss or due to caring for their families. A thousand respondents

were shown fictional résumés and asked their perceptions of the hypothetical applicants. Results showed that job applicants who opted out for family reasons were penalized in that they were seen as less committed to their work, less reliable, and less deserving of being hired, compared to applicants who were unemployed due to job loss. These results also varied by gender: opt-out fathers were penalized more severely than mothers. Weisshaar's study shows how a gap in employment is perceived differentially by gender, thus providing evidence of differential treatment in the labor market.

Experimental methodology has also been extended to the topic of education inequality. Rachel Fish (2016) conducted a survey experiment to assess the role of discrimination in teachers' recommendations for exceptionality testing, a precursor to placement in either special education or gifted/talented programs. In her study, seventy teachers read case studies of male students who varied by race/ethnicity, English language learner status, and their exceptionality characteristics (including their academic, social, and emotional strengths and challenges). After reading the case studies, teachers were asked whether they would refer the student for exceptionality testing. Pager found that black and Latino boys were more likely to be recommended for exceptionality testing when they exhibited behavioral challenges while white boys were more likely to be recommended when they exhibited academic challenges. Her findings suggest that teachers enact differential types of interventions for black and white students, thus providing empirical evidence for differential treatment in education. The present study falls alongside these studies, applying experimental methodology to the question of discriminatory decision-making in school discipline scenarios.

For this study, participants are presented with three fictional vignettes in which a student exhibited a negative behavior at school (see example below). They are then asked to recommend

a disciplinary response for the student in each vignette. Vignettes were designed to fall within the realm of "discretionary incidents" (Fabelo et al 2011), meaning they were not substantial enough to warrant automatic expulsion or suspension. Instead, behaviors could result in a harsher or more lenient punishment based on the inclination of the respondent. Vignettes were written such that the age, socioeconomic status, attendance record, previous discipline record, and family background of the student are unspecified, while the race and gender of the student are signified by the student's name. (See further details on the use of names below.)

> Sample Vignette
> At lunch in the cafeteria, the School Resource Officer saw [Hilary] and another student stand up, get in each other's faces, and shout threats and insults at one another. Before the School Resource Officer could stop the fight, [Hilary] had pushed the other student to the ground and punched her in the face resulting in a bloody nose. After receiving statements from both students, it seems that the confrontation was over a shared romantic interest.

The participants are also given a mock school discipline policy to frame their disciplinary recommendations. Two policy versions, one permitting a high degree of discretion and the other permitting a low degree, were modeled after codes of conduct currently being used in public middle schools. After reading each scenario, participants are asked to respond with their recommended disciplinary action.

> Mock Discipline Policy – High Discretion
>
> In this middle school, principals and teachers have full authority as provided by law to establish and enforce standards and rules as are necessary to create orderly schools and classrooms. According to the Student Code of Conduct, disciplinary consequences may include, but are not limited to, the following:
>
> 1. Detention
> 2. In-school suspension
> 3. Out-of-school suspension
> 4. Expulsion

Mock Discipline Policy – Low Discretion

In this middle school, principals and teachers have full authority as provided by law to establish and enforce standards and rules as are necessary to create orderly schools and classrooms. According to the Student Code of Conduct, disciplinary consequences are assigned to students according to their offense, as outlined below:

1. Defiance or disrespect → detention
2. Inappropriate language → detention
3. Disruption → in-school suspension for one day
4. Leaving school without permission → in-school suspension for one day
5. Skipping → in-school suspension for one day
6. Communicating threats → out-of-school suspension for three days
7. Physical aggression toward another student → out-of-school suspension for four days
8. Assault of school employee → out-of-school suspension for five days
9. Possession of drugs: first offense → out-of-school suspension for seven days
   Possession of drugs: second offense → expulsion

The survey also includes a few brief questions to determine the participants' rationales in assigning their chosen disciplinary action for each case (see list below):

1. Delivering a sufficient punishment for the student's misbehavior
2. Preventing future misbehavior from the student
3. Discouraging other students from similar behavior
4. Minimizing disruption to the learning environment
5. Ensuring the safety of other students
6. Ensuring the safety of school staff
7. Following the school discipline policy

Rationales were designed to address a range of potential decision-making considerations that research shows may vary based on the race and gender of the student. These include educators' interests in preventing disruption, minimizing threats, encouraging "good" behavior, demonstrating punitiveness, and adhering to the school's discipline policy (Morris, 2005; Morris, 2016; Ferguson, 2000). The survey concludes with a short demographic section with questions on the participant's race, gender, age, and profession within the school district (Appendix C). Additional details about the survey are included below in the Variables and Analysis section.

Variables and Analysis

There were three key independent variables for this study: the race of the fictional student in the scenario, the gender of the fictional student, and the degree of discretion permitted by the mock discipline policy. Race and gender were indicated by student name, using methodology designed by Michael Gaddis (2015) in his audit study of racial differences in hiring practices and by Rachel Fish (2016) in her survey experiment on racial differences in student exceptionalities testing. For this study, it was critical that participants did not become aware that the study is assessing racial or gender disparities. The use of names was selected to signify race and gender instead of other possible cues such as facial images or simply stating the student's race because it seemed less likely to trigger the participants' awareness of the racialized nature of the research questions. Three names were selected to represent each of four experimental categories: black male, black female, white male, and white female[1]. To select these names, I reviewed the names used in Michael Gaddis's 2017 study testing race and class associations with names. Using his data, I selected three names for each race-gender combination that had 90% or higher congruent perception rates for race and that had similar proportions of college-educated mothers (between 29-41% of children with these names were born to mothers who had attended college). The names that were selected were Hilary, Susan, and Amy to signal white femininity; Ebony, Kenya, and Tyra to signal black femininity; Cody, Dustin, and Steven to signal white masculinity; and Jamal, Terrell, and Tremayne to signal black masculinity. A pretest was conducted using Amazon Mechanical Turk (MTurk) to determine to what degree participants associated each of the selected names with the race it was intended to signal (see names pretest

[1] While the inclusion of other racial and ethnic categories would be beneficial for future studies, this paper is limited in scope to white and black students, allowing for an investigation into the unique dynamics between these two racial groups which represent the country's dominant racial group and the racial group that receives the harshest disciplinary outcomes in the U.S. (Fenning and Rose 2007; Losen and Gillespie 2012).

survey in Appendix E). The congruence rates between the participants' selected race and the intended race ranged from 87-98% across these twelve names (N=460).

The third key independent variable was the degree of discretion permitted to the participant when making a discipline decision. As described above, mock discipline policies were provided to participants to frame their decision-making, stated in either specific terms (low discretion) or vague terms (high discretion). This was designed to allow for an analysis of whether the degree of discretion permitted by the policy had an impact on how harsh a punishment the participant recommended. For this experiment, there was not be a control condition due to the fact that the study was meant to assess the effect of the *level* of discretion permitted by the discipline policy rather than the effect of having a discipline policy at all. Draft discipline policies are included in the survey document in Appendix C.

Before distributing the survey to participants, it was pretested again using Amazon Mechanical Turk, this time distributing the full survey to 240 MTurk workers in order to assess the effectiveness of the various components of the survey (see MTurk pretest survey in Appendix D). The responses revealed no substantial concerns or areas of confusion related to the survey design. After gathering the pretest data, I thought to add a seventh rationale question asking participants to rate the importance of "following the school discipline policy" in shaping their disciplinary recommendations. The survey was then distributed to the intended sample by means of social media distribution (see details about distribution below).

Each participant was given a series of the same three mock discipline scenarios, though the order of the scenarios varied randomly across participants. Each participant received a single, randomly assigned student name that was held constant throughout their survey. This "split ballot design" was used in order to minimize the degree to which the participant realized the

racialized nature of the research questions (Pager and Quillian 2005). The dependent variable for the study was the type of disciplinary action participants assign to the student in the vignette. After reading each vignette, participants were asked which form of discipline they would recommend for the student among the following options: detention, in-school suspension, out-of-school suspension, and expulsion. If the participant selected either form of suspension, they were then asked how many days of suspension they would recommend using a slider tool ranging from 1-7 days. Next, the survey included a section asking participants to rate the importance of several possible rationales in shaping their disciplinary recommendations. This section of the survey was presented after each vignette without permitting the participant to go back and revise their disciplinary reaction.

Distribution

The target population for this study is current public school educators who work in any position in a public school or district office. Research shows that a variety of school staff are involved in responding to student behavior infractions from start to finish (Lewis and Diamond 2015; Ferguson 2000), such as teachers, administrators, and other school staff who may report student misbehavior, recommend disciplinary actions, decide on the discipline to be given, or enact the disciplinary decision. Survey distribution occurred through the online distribution of a Qualtrics survey, allowing for responses to be completely anonymous. Participants for the survey's primary distribution outlet (via social media) were not compensated for their time, although MTurk pretest participants for both the names pretest and survey pretest were compensated based on the hourly minimum wage in the state of North Carolina.

I went through several stages to find a distribution outlet before ultimately using social media-based. The initial plan for survey distribution was to identify one or more public school districts that would permit me to distribute the survey to their full staff via email. To do so, this required the permission and cooperation of district-level administrators, a requirement that proved extremely difficult to meet. After pursuing survey distribution opportunities in fourteen school districts across five states, I was unable to find any districts willing to distribute the survey within the timeline necessary to complete this research. The second plan was to distribute the survey to the alumni network of a university's school of education by means of their email list-serve. This required the permission and cooperation of the school's administrators, and this too was something I was unable to find. The third plan was distribution via social media, which was the distribution outlet I ultimately used.

The survey was distributed by means of snowball sampling through the researcher's personal and professional social media platforms, as well as by "cold-calling" (or in this case, "cold-messaging") a variety of well-known educators whose social media pages had sizeable followings. The link to the Qualtrics survey was posted on the researcher's own Facebook page and Twitter feed and was sent out via email to a variety of personal and professional contacts. These contacts were asked to respond to the survey if they were a current public school educator and to forward the link to any public school educators they knew. In order to ensure participants were over 18 and current public school educators, two qualifying questions were included in the survey asking participants to indicate "yes" or "no" to both factors. If the respondent selected "no" for either question, they were thanked for their time and the survey was terminated.

RESULTS

Quantitative analysis of the survey responses was conducted using StataSE 15 software to test the following hypotheses:

Hypothesis 1: Participants will prescribe harsher discipline to black students than to white students and to boys than to girls, with black boys receiving the harshest punishment, followed by black girls, white boys, and white girls.

Hypothesis 2: When participants are allowed a higher degree of discretion, their disciplinary decisions will show heightened rates of disparity within the sequence described in Hypothesis 1.

Hypothesis 3: Respondents will tend to punish white boys and girls with rationales that involve preventing future misbehavior and restoring good behavior. Respondents will tend to punish black boys and girls with rationales that involve protecting other students and staff and minimizing disruptiveness.

Multivariate analysis of participants' disciplinary decisions was conducted using four possible operationalizations of discipline severity: 1) the type of discipline recommended, 2) whether the student was punished, 3) whether the student received exclusionary discipline, and, 4) in the case that suspension was given, the number of days of suspension. For this study the unit of analysis is the scenario rather than the person, resulting in a consideration of the unique characteristics of each disciplinary incident and its contextual factors that may shape disciplinary outcomes. I also include an analysis of the MTurk survey pretest results as an additional data

source. Analyses of both samples are included here and are presented in sequence, allowing for an examination of the variations in findings across both samples.

Results from Educator Sample

In this section I describe the descriptive and multivariate results of the survey experiment, beginning with an overview of the participants' demographic characteristics and the scenarios to which they responded (see Table 1). Responses were dropped if the survey duration fell in the bottom fifth percentile, representing a completion time of 3.5 minutes or less, in order to remove cases in which the participants did not thoroughly consider the information presented. This resulted in a total 243 participants. The majority of participants in the sample were white (84.5%), female (88.7%), and in the 30-39 and 40-49 age brackets (35.3% and 28.3% respectively). Most participants were employed as public school teachers (78.8%). The three key independent variables for the study (student race, student gender, and discretion level) were varied randomly among the participants and are thus divided fairly evenly across the discipline scenarios (Table 2). Given that all three key independent variables are binary, there are a total of eight possible experimental conditions. Of the data gathered, the smallest sample size for a single experimental condition was 84 while the largest was 123.

A descriptive overview of the participants' disciplinary recommendations is shown in Table 3. Only a small minority of scenarios resulted in participants recommending "none of the above" disciplinary actions (8.7%). These relatively low percentages follow logically from the fact that the incidents were written to intentionally suggest a level of misbehavior that would warrant "serious" punishment including suspension and expulsion. Among those scenarios in which detention, suspension, or expulsion were selected, the least common disciplinary response

was detention (4.8%), followed by expulsion (12.4%). Out-of-school suspension was the most common disciplinary response (47.1%), followed by in-school suspension (27.0%). Participants' ratings of the degree to which the seven rationales influenced their decision-making were fairly high (Table 4). On a scale of one to five, the average ratings fell between 3.56 and 4.15 across the six rationales, all noticeably higher than the 2.5 halfway point. Using participants' ratings of these seven rationales, I conducted factor analysis in order to create a smaller number of simplified variables for use in the regression analysis (Table 5). The factor analysis resulted in two dimensions of disciplinarity, which I term "authoritarianism" and "protectiveness."

Multivariate ordered regression, clustered by participant, was conducted to assess the combined and interactive effects of race, gender, and discretion while also controlling for a variety of additional variables. The control variables for this study include the nature of the incident presented to the participant (Appendix C), the race and gender of the participant (Table 1), and the disciplinary factors created from the rationale responses (Table 5). The key dependent variable is the type of discipline recommended (detention, in-school suspension, out-of-school suspension, or expulsion), operationalized as an ordinal variable.

The results of multivariate ordered regression analysis are shown in Table 6. In Model 1, the effect of student race (whiteness as compared to blackness) and student gender (masculinity as compared to femininity) are tested. The effect of whiteness has a very small, positive value (0.086) and is not significant at the 0.1 level, while the effect of masculinity has a very small, negative effect (-0.117) but is not significant at the 0.1 level either. In the second model, student race and gender are interacted. As with the first model, the race and gender effects are not significant at the 0.1 level and appear very close to 0 in magnitude. Model 3 introduces the incident type and the level of discretion provided in the mock discipline policy. These new

variables are all significant at the 0.01 level and are larger in magnitude than the effect of student race and gender (0.360 for a fight-related incident, 1.085 for a marijuana-related incident, and -1.050 for a high level of discretion). The variables for both the fight-related incident and the marijuana-related incident are positive compared to the defiance scenario, while high discretion has a negative effect on discipline severity. In Model 4, the incident type and discretion level are interacted. All three variables remain significant at the 0.01 level except for the interaction of the fight-related incident and high discretion. All three effects are fairly sizable in magnitude. For both Model 3 and Model 4, the effect of student race and gender remain insignificant and small in magnitude. In Model 5, variables are added for the race and gender of the participant, which are then interacted in Model 6. Neither variable, nor their interaction, are significant at the 0.1 level. When they are added to the model, the incident types and discretion levels that were significant in Model 4 remain significant, while student race and gender remain insignificant. Two factor variables for the six decision-making rationales are added to the analysis in Model 7. Authoritarianism is significant at the 0.01 level but protectiveness is not significant. In Model 8 and 9, the rationale factors are interacted with the race and gender of the participant and with each other. Model 8 includes student race and gender as independent variables while in Model 9 they are interacted. None of the interactions of the rationale factors with participant race and gender are significant, but the effect of authoritarianism remains significant in both models. Student race and gender fail to show statistical significance regardless of whether they are included as independent or as interacted variables.

Additional multivariate analyses were conducted using three additional operationalizations of discipline severity, including the same sequence of independent variables. These analyses include logistic regression predicting whether the student was punished (Table

7), logistic regression predicting whether the student received exclusionary discipline (Table 8), and regression analysis of the number of days of suspension (Table 9). As with the previous analysis, student race and gender remain statistically insignificant in these additional operationalizations of discipline severity.

Coefficient plots were also created for all four analyses to visually assess effect sizes across the models, focusing on models 2, 4, 6, and 9 which include the full range of intersections. While the values of these coefficients are statistically insignificant, it is potentially informative to look at the trends in effect sizes across the models. The coefficient plot of the log odds predicting the type of discipline recommended (Figure 1) shows that, while the effect sizes of the incident type, discretion level, rationale factors, and participant characteristics are somewhat large in magnitude, the effect of student race and student gender remain close to zero across all the models. The effect of white femininity and white masculinity appear slightly positive, suggesting that being a white boy or girl may increase the likelihood of punishment. The effect of black masculinity has a slight negative value, suggesting that being a black boy may decrease the likelihood of punishment. In the model depicting likelihood of punishment (Figure 2), the effects of student race and gender are nearly indistinguishable from zero. The coefficient plot of the likelihood of exclusionary discipline (Figure 3) shows similar trends for the effects of white femininity, white masculinity, and black masculinity as were shown in the coefficient plot of the type of discipline recommended (Figure 1). Finally, in the coefficient plot of days of suspension (Figure 4), the effect sizes of white femininity is similar to those in the analyses of discipline type and likelihood of exclusionary discipline, while white masculinity now has a slight negative value and black masculinity has a slight positive value. Despite these variations in magnitude for the effect sizes of student race and gender, it is important to keep in mind that, across all four

analyses, none of these effects are statistically significant and, in fact, their values fall in contrast to those predicted by the hypotheses.

In sum, the results of multivariate regression analysis of these survey responses fail to confirm the study's hypotheses. Considering that the sample for these analyses was collected by means of social media distribution, there exists the possibility that the null findings are due to the nature of the sample. Next, I provide an analysis of the results from the MTurk pretest survey in order to assess whether a sample of anonymous, non-educators produces different results.

Results from MTurk Pretest

Although the MTurk pretest was initially designed to assess the effectiveness of the survey itself, it presents a useful dataset for comparison in that a) the respondents were recruited from a very different source, b) the sample size is substantial, and c) the surveys distributed to the two groups were nearly the same. Two changes were made between the distribution of the pretest and the distribution of the survey to educators. First, I added a question to collect information about the participants' jobs in their school district in order to assess the range of educational staff members present in the sample. Second, I thought to add a seventh rationale to assess the degree to which respondents felt compelled to adhere to the discipline policy that was given. Since these two survey questions were not included in the MTurk pretest, they are not available for analysis using this sample. As with the sample of educators, I dropped responses from participants whose response time was in the lowest fifth percentile, representing a completion time of 2.5 minutes or less. Additionally, one respondent was dropped who identified as neither male nor female, given that with only one respondent in this category the sample

lacked sufficient power to run the models with third gender category as a variable. This left a sample size of 239 participants who responded to 717 scenarios.

As with the respondents in the educator sample, the participants in the MTurk sample were primarily white (78.2%) but they were fairly balanced on gender (42.3% men and 57.7% women) and were younger in age (38.5% were younger than 30 and 39.3% were between 30-39 years old) (Table 10). Across the study's eight experimental conditions, the sample sizes were slightly smaller than in the educator sample, with 84 as the smallest sample size and 102 as the largest (Table 11). Descriptive results of the participants' disciplinary recommendations (Table 12) show that participants were least likely to recommend no punishment (2.1%) followed by detention (8.4%) and expulsion (18%). The majority of the respondents recommended suspension, with 23.2% of scenarios resulting in in-school suspension while 48.4% scenarios resulted in out-of-school suspension. Participants' ratings of the importance of the six rationales (Table 13) ranged from an average of 3.66 for "ensuring the safety of staff" to an average of 4.09 for "discouraging other students from similar behavior." As with the educator sample, I conducted factor analysis in order to create a smaller number of simplified variables for use in the regression analysis (Table 14). Factor analysis resulted in one factor, which I term "disciplinary zeal."

The same set of analyses were run on the MTurk sample as were run on the educator sample, including multivariate regression analysis predicting the severity of the disciplinary response (Table 15), the likelihood of punishment (Table 16), the likelihood of exclusionary discipline (Table 17), and the number of days of suspension in the case that suspension was given (Table 18). The same sequence of variables was used for all four analyses, resulting in the same sequence of nine models.

In the first model of the analysis predicting disciplinary response (Table 15), the independent variables are insignificant and their effects are fairly small in magnitude (0.151 for whiteness and 0.039 for masculinity). In Model 2, student race and gender are interacted. Their effects remain small in magnitude and insignificant. Model 3 includes variables for the incident type as well as the level of discretion. Both the fight-related effect and the marijuana-related effect are positive and significant at the 0.01 level compared to the excluded category which was the defiance-related incident. The effect of high discretion is also significant but has a negative relationship with discipline severity, suggesting high discretion tends to decrease the severity of participants' recommended disciplinary responses. When the incident type and discretion level are interacted in Model 4, their effects remain significant except for the effect of high discretion in defiance-related incidents. The effects of student race and gender remain insignificant in both Models 3 and 4. Participant characteristics are added to the analysis in Models 5 and 6, first assessing their independent effects and then their interacted effects. When considered independently, the effect of masculinity is significant at the 0.05 level; however, when considered as interacted variables, participant masculinity becomes insignificant. In Model 7, a factor variable for "disciplinary zeal" is added to the analysis. It is has a statistically significant, positive effect on discipline severity. In Model 8, disciplinary zeal is interacted with participant race and gender. Only the effects of disciplinary zeal for non-white women remains significant. Model 8 includes the full set of control variables with participant race and gender operating independently, while Model 9 assesses their interacted effects. Across all nine models, the effect of student race and gender on discipline severity remain insignificant.

The same three additional multivariate analyses were conducted for the MTurk sample as were conducted for the educator sample, in order to assess the effects of these same independent

variables on other operationalizations of discipline severity. Student race and gender remain statistically insignificant in the analysis predicting days of suspension (Table 18). However, in the analyses predicting the likelihood of punishment (Table 16), black masculinity has a significant negative effect on the likelihood of punishment while white masculinity has a significant positive effect compared to black femininity. This suggests that being a white boy increases the likelihood of punishment while being a black boy decreases it. A similar trend is found in the regression predicting the likelihood of exclusionary discipline (Table 17): being a black boy seems to decrease a student's likelihood of punishment while being a white boy seems to increase it. The effect sizes for models 2, 4, 6, and 9 across all four analyses are depicted in coefficient plots (Figures 5-8). While the results of the MTurk survey yield several instances in which student race and gender are significant, their relationship to discipline severity is in the opposite direction of that predicted by the hypotheses.

In sum, the results of multivariate regression analysis of both the educator sample and the MTurk sample fail to confirm the study's hypotheses. In almost all analyses, student race and gender have no significant effect on discipline severity. In the two analyses in which student race and gender do significantly affect discipline outcomes, participants are more lenient toward black boys and harsher toward white boys, standing in contrast to the study's hypotheses. In the next section I will assess these unexpected findings in relation to the previous sociological literature, offering three possible explanations for their divergence.

DISCUSSION

At first encounter, these findings seem to suggest that differential or discriminatory treatment does not explain the pronounced and persistent patterns of school discipline disparity across the United States (Losen and Gillespie 2012; Fabelo et al. 2011; Fenning and Rose 2007). After considering discipline severity in terms of the types of discipline recommended, the likelihood of punishment, the likelihood of exclusionary discipline, and the number of days of suspension, my findings provide essentially no evidence for the discriminatory treatment of black students or the preferential treatment of white students. Instead, the data seem to suggest that participants recommended virtually the same degree of severity for black and white students. This finding is at odds with much of the existing literature on racial disparities in discipline. What do these findings tell us?

First to consider is the possibility that the survey results are an accurate depiction of real-world disciplinary decision-making, suggesting that school discipline disparities are not driven by discriminatory treatment. Second, I consider the possibility that the study's sample was non-representative in that it was oversaturated with people who are disinclined to demonstrate racial discrimination. The third possible explanation is that the participants were influenced by social desirability bias such that, even if they did have racial prejudice, the pressure to appear "non-racist" swayed their answers. Finally, I explore the possibility that the survey vignette methodology itself is a flawed approach to addressing the research question; perhaps when presented with hypothetical discipline scenarios, people's decision-making does not align with

the ways they behave in an actual school context. I will review each of these explanations in further detail in the sections that follow.

Explanation One: Students are Truly Treated Equally

It is critical to scientifically consider all possible causes of school discipline disparity, turning one's attention to the possibility that discriminatory decision-making may *not* play a part. As discussed in the literature review above, there is an enduring biological determinism argument suggesting that black students are inherently more prone to deviant behavior, and it is thus their differential behavior instead of differential treatment that results in their higher rates of punishment (Skiba et al. 2002).

Previous research has documented only limited evidence for differential rates of misbehavior between black and white students (Roque 2010, Skiba et al 2002; Welch and Payne 2010; Rocque and Paternoster 2011). What little evidence scholars do find is overshadowed by more robust evidence for differential treatment (Skiba et al. 2002; Skiba et al. 2011; Skiba et al. 2014; Roque 2010). Scholars have, for example, refuted the hypothesis that black students' higher rates of punishment correspond with more serious types of behavioral infractions. Rather, studies find that black students are referred at higher rates for behaviors that are subjectively interpreted, and it is this disparity, rather than a disparity in the gravity of infractions, that constitutes the bulk of discipline disproportionality (Skiba et al 2002). Others tested whether teachers' assessments of their students' tendencies to "act out" could account for the discipline gap and found that these assessments, intended as a measure of differential behavior, failed to provide sufficient explanatory power for the breadth of the discipline gap (Rocque 2010). Furthermore, there is evidence that higher rates of exclusionary punishment align with a higher

proportion of black students in the student body, regardless of students' actual rates of misbehavior (Welch and Payne 2010; Rocque and Paternoster 2011). In short, quantitative evidence for racially disparate rates of misbehavior is overshadowed by more conclusive evidence of differential treatment. Ethnographic studies provide a more nuanced look into the comparative strength of the differential treatment and differential behavior hypotheses using data from researchers' observations in school settings and their interviews with students and staff. These scholars find strong evidence for differential treatment, whereby school staff select black students for punishment at higher rates and afford them harsher punishments for the same behaviors white students portray (Ferguson 2000; Lewis and Diamond 2015; Morris 2016; Morris 2005).

Amidst this substantial evidence for differential disciplinary treatment, my finding that, when responding to survey vignettes, participants do *not* recommend harsher punishments for black students is perplexing. Next I assess the nature of the sample itself as a possible explanation for the study's null findings.

Explanation Two: These Aren't the Racist Ones

Next to consider is the possibility that sampling bias skewed the results such that the findings from the sample of educators don't reflect the true nature of the population of U.S. public K-12 educators (Singleton and Straits 2018). Given that the data were collected by means of social-media-based snowball sampling that originated from the researcher's personal contacts, one could imagine that this group of respondents would be less prone to racial prejudice given my own anti-racist convictions and my gravitation toward others with similar racial justice attitudes. Research shows that social networks tend to be politically and socioeconomically

homogenous (McPherson et al 2001); thus, it is probable that the study's respondents, recruited by means of my personal Facebook page, Twitter feed, alumni networks, and friend networks, are more prone to anti-racist ideologies than the average American educator[2]. If so, one could argue that, although Americans educators tend to make racially discriminatory discipline recommendations, those weren't the ones who were captured in my sample. In other words, one might argue, "these aren't the racist ones."

Arguably, the best evaluation of this explanation is done empirically with a comparison of the results from the social media and MTurk samples. If one were to argue that the null findings in the social media sample likely result from the non-prejudicial characteristics of the researcher's personal networks, it would be unexpected to find the same null findings in a sample of MTurk participants who are presumably no less likely to be prejudiced than the average American. Previous research on MTurk survey-takers shows that, while MTurk participants do tend to be slightly younger and more educated than the average American, they are nevertheless more representative than other forms of convenience samples often used in social science experiments (Gaddis 2017; Berinsky, Huber, and Lenz 2012). In fact, when comparing demographics between this study's educator sample and the MTurk sample, we find that the MTurk participants are on average younger and more racially diverse than the social media sample, and they include substantially more men among their ranks. Despite these differences, analysis of MTurk participants' disciplinary recommendations are similar to the results of the educator sample in that they fail to show any statistically significant relationships with student race or gender occurring in a direction that would explain the discipline gap. Even this sample of anonymous, online strangers produces no evidence of discriminatory decision-making as a cause

[2] Or, perhaps selection bias is to blame for the lack of evidence for racial discrimination. Perhaps, despite my intentions to not advertise the study's racial elements, participants opted in if they were concerned about racial disparities in school discipline.

of discipline disparity. Therefore, it seems unlikely that the null findings in the social media sample are due to these participants' higher proclivity for anti-racist ideologies.

Next, I consider whether the null findings could be due to social desirability bias such that participants, regardless of any anti-racist leanings, knew to obscure their potentially prejudiced attitudes.

Explanation Three: They Knew the "Right" Answers

While the first explanation proposes that the sample of participants is the underlying cause of the study's unexpectedly null findings, this second explanation proposes that the participants, regardless of their political stances, would know it is not socially desirable to show preferential treatment of white students or discriminatory treatment of black students. This tendency could explain the divergence of my findings from the consistent patterns of racially discriminatory treatment that emerge from empirical studies (Skiba et al. 2002; Skiba et al. 2011; Skiba et al. 2014; Roque 2010).

When influenced by social desirability bias, research participants attempt to steer their answers toward those that demonstrate socially desirable traits and attitudes (Pager and Quillian 2005; Singleton and Straits 2018:598). In a conscious or unconscious effort to bolster their self-esteem or make a good impression, participants shape their responses to appear healthy, happy, mentally fit, and free of racial prejudice, even if doing so means departing from the truth (Pager and Quillian 2005; Singleton and Straits 2018:341)[3]. Split-ballot survey designs, such as that

[3] Compelling evidence for social desirability bias emerges from Derek L. Phillips and Kevin J. Clancy's study (1970) that find that middle-class participants self-report greater happiness, mental health, and lower racial prejudice compared with working-class participants. Similarly, Geoffrey T. Wadtke (2016), in a study titled "Are Smart People Less Racist?" finds similar trends when comparing participants' higher cognitive abilities compared with those with lower cognitive abilities. These scholars conclude that, rather than signifying that these groups are truly "better off" in these ways, wealthier and "smarter" participants are simply more concerned with their appearances

used in the present study, are often used to avoid this form of bias, and are typically found effective (Pager and Quillian 2005). However, in the case of race-related studies, participants may notice that race has been signaled and therefore consciously or unconsciously adjust their behavior to mask any negative reaction they may have toward a black student or a positive reaction toward a white student.

The social desirability explanation is additionally meaningful given the nature of current American racial ideology. Since the late 1900s, a historical shift occurred whereby white Americans began to disavow their earlier overt support for racial segregation and discrimination and to embrace, at least in principle, racial equality (Bonilla-Silva 2018; Omi and Winant 2014; Dixon 2017; Pager and Quillian 2005). As of the late 20th century and into today, a form of "new racism" has gained popularity that is more covert and sophisticated that previous overt forms of "old racism" (Bonilla-Silva 2018; Pager and Quillian 2005). As a result, it has become rare to find white Americans who voice support for previously popular forms of racial discrimination and segregation, such that a number of questions were "retired" from national attitude surveys (Dixon 2017: 93). This rosy image is nevertheless confounded by the troubling endurance of racial inequality in American society in domains as far-reaching as education, housing, health, employment, politics, and the criminal justice system (Bonilla-Silva 2018; Wodtke 2016; Ioanide 2015; Sears et al 2000; Sears 1988; Omi and Winant 2014; Bobo 1988).

Political psychologists provide empirical support for this apparent contradiction, which these scholars term the "Principle-Implementation Gap" or the "Principle-Policy Paradox." Results from attitudinal studies show that, although white survey respondents signal their support

---

(Phillips and Clancy 1970; Wadtke 2016). Thus, they may have consciously or unconsciously adapted their behavior from what would otherwise be their "true" response.

for racial equality in principle, they oppose the policies that would be required for its implementation, (Dixon et al 2017; Wodtke 2016). Social psychologists come to similar conclusions, finding that although white Americans tend to renounce racially prejudiced principles in self-administered self-reports, they more readily express racial prejudice in other contexts (e.g. when the interviewer is also white or when a lie detector device is used). Furthermore, studies reveal that even when participants are sincerely unaware of having racial prejudice, it nevertheless emerges in tests of implicit or unconscious racial bias (Sears et al 2000; Eberhardt 2005; Wodtke 2016; Ioanide 2015). Some examples from the field of social psychology include experiments that used timed exercises to test for the relative accessibility of positive versus negative associations and experiments that assess participants' subconscious cognition by tracking their eye movements (Sears et al 2000; Gilliam et al 2016). These various forms of implicit bias can operate in spite of consciously held attitudes and may persist in influencing behaviors (Pager and Quillian 2005). In short, while most white Americans comprehend that overt racism is not considered socially acceptable, they more readily demonstrate racial prejudice in studies that allow for its covert expression (Eberhardt 2005; Ioanide 2015).

In the field of sociology, this new form of racial ideology has been explained by the theories of "symbolic racism" (Sears 1988), "laissez-faire racism" (Bobo, Kluegel, and Smith 1997), and "colorblind racism" (Bonilla-Silva 2018). While these theories have important distinctions, all of them propose that in this new era, overt racism has essentially "gone underground." Racism is now heard in different tones, emphasizing at times the values of liberalism and group interests in ways that place the blame for racial inequality on black people's own moral failures. David Sears explains that these new forms of racism "blended some anti-

black feeling with the finest and proudest of traditional American values": hard work, individualism, and reward by merit. (1998:54). White Americans now use "code words" to explain what previously would have been overtly prejudiced viewpoints, using instead terms like "law and order," "forced busing," and "states' rights" (Omi and Winant 2018: ch. 7-8). In short, scholars across the disciplines of sociology, psychology, and politics conclude that, in simple terms, "prejudiced people are not the only racists in America" (Wellman 1993:27). The recent shift to laissez-faire, symbolic, or color-blind racial ideologies has resulted in an American public that is highly aware that it is distasteful to be overtly racist, and has become highly skilled at the cloaked expression of their racially prejudiced ideals. In this context, it is all the more likely that social desirability would prompt participants to adjust their answers in a way that may produce this study's puzzling null findings.

This explanation, in light of the literature, seems probable; however, it rests on the participants having deduced, consciously or subconsciously, the study's intent to study racial trends in their responses despite the split-ballot design. The fourth explanation does not rest on this prerequisite, nor does it require any uniquely "non-prejudicial" characteristics of the sampled participants. Instead, the fourth explanation posits that it is the very nature of survey experiments, specifically in the case of racialized discipline-related decisions, that swayed participant responses from the way educators behave in real-world scenarios.

Explanation Four: Talking the Talk but Not Walking the Walk

The previous explanation proposes that, due to social desirability bias in the context of "new racism," participants adapted their answers away from their true attitudes. In contrast, this fourth explanations suggests a more sympathetic stance, suggesting that participants may truly

believe they are responding as they would in an actual disciplinary incident; however, the very nature of hypothetical survey questions differs substantially from the actual disciplinary decision-making context such that they react substantially differently. Pager and Quillian (2005) explain how vignette studies rest on the assumption that participants' self-reported responses to hypothetical scenarios accurately mirror how they would behave in the actual situation. This is called attitude-behavior correspondence, and suggests that people's behavior is a response to conscious thoughts and decisions. While some psychological studies find that conscious attitudes do in some ways influence behavior, others report zero or negative correlation between the two. There is enormous variation in the literature, and the relationship is far from clear-cut.

Survey vignette methodology was first developed for use in measuring racial attitudes in the 1970's and has since been expanded upon for a broad variety of tests of racial discrimination (Pager and Quillian 2005). Survey experiments have been used to provide evidence for discriminatory decision-making in the realms of hiring decisions (Pager 2003; Gaddis 2015), housing selection (Flage 2018; Gaddis and Ghoshal 2015), and student exceptionality testing (Fish 2016), several of which were reviewed in this study's research methods section. In these three cases, participants' hypothetical reactions to fictional vignettes appear to provide evidence for the types of racially discriminatory incidents that play out in the "real world." Meanwhile, alongside these types of decision-making contexts is another set of contexts that don't seem to align as closely with the empirically demonstrated patterns of disparities.

One such study was conducted by Walter Gilliam and his colleagues (2016), who asked early childhood educators to identify challenging behaviors in a video of preschoolers despite the fact that none were included in the video. When they tracked the educators' eye gaze, they found that participants were more keenly focused on black students and particularly black boys when

watching for problematic behavior. However, when the same educators were shown a vignette of a student misbehaving and asked how likely they were to recommend suspension or expulsion, they found no effect of student race. Other curious null findings emerge from studies in which participants are asked to appraise the seriousness of police use of excessive force (Rome 1995), the vulnerability of rape survivors, (Donovan 2007), and the perceived pain levels of hospital patients (Haider et al 2011), none of which find significant effects in terms of the race of the person in the experimental vignettes.

Devah Pager and Lincoln Quillian (2005) discuss this disjuncture in their study of employers' hiring behavior. To do so, they conducted a survey experiment and an audit study, both of which assessed employers' likelihood of hiring a black or white ex-offender for an open entry-level position. No matter how they analyzed the data, they found drastically different results between the two studies: while in the survey experiment, employers professed a willingness to hire both black and white ex-offenders, their actual hiring behavior showed they were much less likely to hire a black ex-offender than a white ex-offender. The paper reveals the limitations in using experimental survey results to draw conclusions about discriminatory hiring practices. As the authors state, "Accepting the survey results as an accurate indicator of the opportunities available to blacks and ex-offenders would grossly understate the barriers to employment they face" (364).

Pager and Quillian (2005) present a number of psychological and contextual factors that diverge significantly between a hypothetical scenario and an actual incident. For example, they state, "The intrapsychic processes that promote discrimination are likely to be more strongly activated in the context of a live interaction than in the abstract context of a survey question" (359). Evidence shows that when whites interact directly with blacks they show higher

levels of discomfort, which can result in discriminatory behavior despite their avowed non-prejudicial racial attitudes. Discomfort in this case is measured in terms of participants' speech errors, the number of questions they ask, and their haste in terminating an interviewer when they are paired with a black conversation partner, behaviors that would not be captured in a survey experiment (359). There are also a variety of situational factors that may explain this disjuncture. Discriminatory behaviors are context-specific and are shaped by the complex interplay of social norms that emerge from an actual situation. For example, participants may use different criteria in evaluating a hypothetical scenario than they do in an actual scenario (370), especially when the scenario is one that would elicit a more intense array of emotions or a stronger sense of threat when experienced in-person. Pager and Quillian explain, "The live interaction may trigger feelings of fear, anxiety, or threat in ways that a recited vignette does not" (2005:371). This explanation could potentially account for other scholars' null findings in the realm of police violence (Rome 1995), rape (Donovan 2007), and patient pain levels (Haider et al 2011), all of which would likely evoke heightened emotion when encountered face-to-face as compared to the context of a survey vignette.

For these various reasons, Pager and Quillian propose that survey experiments may be more effective in estimating behaviors in contexts that are not dependent on the nature of live interaction, contexts such as voting behavior or consumer behavior (372). While survey data may be able to demonstrate how participants *think* they make decisions or *think* they act, these beliefs seem to have a limited causal role in shaping behaviors in actual decision-making scenarios that are highly charged or emotional in nature. This explanation resonates with the current study's use of discipline scenarios that may likely evoke fear, anger, or a sense of loss of control when experienced in person, but comparatively reduced levels of emotions when read on a screen.

CONCLUSION

This study was designed to use survey vignette methodology to formally test whether educators demonstrate racially discriminatory decision-making when responding to student misbehavior incidents. This represents the first use of experimental methodology in the sociological study of school discipline disparity and the first attempt to isolate and test the potentially causal role of differential treatment in generating the discipline gap. While the results from previous survey experiments suggest this methodology's promising ability to explain a wide range of racial disparities (Pager 2003; Gaddis 2015; Flage 2018; Gaddis and Ghoshal 2015; Fish 2016), this study does not follow suit. Results of multivariate regression analysis suggest that, instead of showing patterns of discriminatory decision-making, participants recommended nearly identical disciplinary responses across race and gender categories. In the context of current quantitative and ethnographic literature supporting differential treatment as an explanation for the discipline gap, these findings are unexpected.

I thus provide several possible explanations, including the possibility that 1) differential treatment does not play a causal role in discipline disparity, 2) the responses from the study's sample of educators are not representative of the population, 3) the participants' responses are biased due to the pressures of social desirability, and 4) the participants' responses to hypothetical vignettes do not reflect their actual decision-making patterns. After considering these four explanations in light of the surrounding sociological literature, it appears likely that this study's null results are due to the scenarios' emotionally charged nature, resulting in

decision-making patterns that do not carry over when the scenarios are read on a screen instead of experienced in the flesh.

There are several limitations to this study, some of which were addressed in the discussion section. In addition to the possible influences of sampling bias and social desirability bias, there is the possibility that participants "opted in" to the study if they were passionate about racial justice, despite the study's careful attempts to obscure the study's racialized research questions. Furthermore, the sample size for this study was somewhat small: if the type of misbehavior is included as a fourth key variable, the sample sizes within each experimental category range from 28 to 41 in the social media sample and 25 to 35 in the MTurk sample. While a larger sample size would have been ideal, the trends presented from the present study, including comparative analyses of the social media sample and the Mturk sample, do not reveal support for the hypotheses, even if the results were statistically significant. Thus, perhaps the ultimate limitation to the study is that presented by Devah Pager and Lincoln Quillian (2005), namely, their suggestion that survey experiments are a flawed methodology for testing discrimination in scenarios that, when experienced in actuality, are stress-ridden and emotionally charged.

This paper opens up a realm of questions for further exploration, including both substantive and methodological concerns. As for the field of research on school discipline, questions remain as to the potentially causal role of discriminatory decision-making. While the present study attempted to provide a direct test of the differential treatment hypothesis, the null results leave these questions unanswered. Methodologically, future studies would be helpful to assess the divergence between survey experiments and audit studies and the contextual factors

that shape the utility of survey experiments. Future research is critical to uncover these many unturned stones.

APPENDIX A: TABLES AND FIGURES FOR EDUCATOR SAMPLE

Table 1. Participant characteristics in educator sample

	Number	Percentage
Race*		
White or Caucasian	239	84.50%
Black or African American	32	11.30%
Hispanic or Latino	8	2.80%
Asian	4	1.40%
Other	11	3.90%
Gender		
Woman	251	88.70%
Man	32	11.30%
Other	0	0.00%
Age		
Younger than 30	48	17.00%
30-39	100	35.30%
40-49	80	28.30%
50-59	39	13.80%
60-69	16	5.70%
70 and older	0	0.00%
Job		
School teacher	223	78.80%
School administrator	27	9.50%
Other school staff	26	9.20%
District administrator	3	1.10%
Other district staff	4	1.40%

Observations (N)=243
*Racial categories were not exclusive, so cumulative tallies will be greater than N.

Table 2. Distribution of independent variables in educator sample

Student Race and Gender

	Black or African American			White or Caucasian		
Discipline policy	Girl	Boy	Total	Girl	Boy	Total
Low discretion	84	99	183	120	111	231
High discretion	102	117	219	93	123	216
Totals	186	216	402	213	234	447

Observations (N) = 849

Table 3. Distribution of participants' recommended disciplinary actions

	Frequency	Percent
None	74	8.7%
Detention	41	4.8%
In-School Suspension	229	27.0%
1-2 days	130	56.8%
3-4 days	72	31.4%
5-7 days	27	11.8%
Out-of-School Suspension	400	47.1%
1-2 days	86	21.5%
3-4 days	166	41.5%
5-7 days	148	37.0%
Expulsion	105	12.4%
N	849	100.0%

Table 4. Educators' average ratings, on a scale of 1-5, of the importance of seven rationales

Rationale	Average rating
Providing sufficient punishment for the student's behavior	3.68
Preventing future misbehavior from the student	3.98
Discouraging other students from similar behavior	3.93
Minimizing disruption to the learning environment	3.80
Ensuring the safety of other students	4.15
Ensuring the safety of staff	3.73
Following the discipline policy	3.56

Observations (N) = 849

Table 5. Factor loadings and unique variances for rationales in educator sample

Variable	Factor 1	Factor 2	Uniqueness
Providing a sufficient punishment for the student's misbehavior	**0.8016**	0.2001	0.3174
Preventing future misbehavior from the student	**0.6119**	0.4220	0.4475
Discouraging other students from similar behavior	**0.7294**	0.3843	0.3203
Ensuring the safety of other students	0.0919	**0.8029**	0.3468
Ensuring the safety of school staff	0.2168	**0.7910**	0.3273
Minimizing disruption to the learning environment	0.2488	**0.7008**	0.4469
Following the school discipline policy	**0.7374**	-0.0545	0.4532

* Those I consider influential, using 0.5 as the cut-off, are marked in bold font. From the factor loading results, we can see that Factor 1 is closely related to the participants' rating of the importance of punitive or authoritarian components, while Factor 2 relates to ensuring the safety of students and staff and creating an effective learning environment. Thus, I will refer to Factor 1 as "Authoritarianism" and Factor 2 as "Protectiveness."

Table 6: Ordered logistic regression predicting disciplinary response, clustered by participant, using educator sample

VARIABLES	Model 1 B	Model 2 B	Model 3 B	Model 4 B	Model 5 B	Model 6 B	Model 7 B	Model 8 B	Model 9 B
Student Characteristics (Ref. = Black, Girl)									
White	0.0862	0.113	0.0140	0.0334	0.0461	0.0470	-0.00616	0.0793	-0.0116
	(0.159)	(0.239)	(0.232)	(0.231)	(0.230)	(0.230)	(0.231)	(0.152)	(0.227)
Male	-0.117	-0.0910	-0.110	-0.110	-0.109	-0.109	-0.124	-0.0777	-0.167
	(0.159)	(0.232)	(0.233)	(0.232)	(0.233)	(0.232)	(0.231)	(0.155)	(0.231)
White X Male		-0.0503	0.0673	0.0605	0.0351	0.0380	0.105		0.171
		(0.319)	(0.312)	(0.312)	(0.311)	(0.311)	(0.307)		(0.308)
Incident Type (Ref. = Defiance)									
Fight			0.360***	0.473***	0.474***	0.475***	0.474***	0.494***	0.495***
			(0.105)	(0.146)	(0.146)	(0.147)	(0.162)	(0.167)	(0.167)
Marijuana			1.085***	1.975***	1.976***	1.977***	1.992***	2.043***	2.045***
			(0.193)	(0.311)	(0.313)	(0.314)	(0.332)	(0.331)	(0.331)
Discretion Level (Ref. = Low)									
High Discretion			-1.050***	-0.586***	-0.604***	-0.604***	-0.684***	-0.666***	-0.669***
			(0.175)	(0.197)	(0.198)	(0.199)	(0.203)	(0.204)	(0.204)
Incident Type X Discretion Level									
Fight X High				-0.209	-0.210	-0.210	-0.263	-0.311	-0.311
				(0.209)	(0.210)	(0.210)	(0.238)	(0.243)	(0.243)
Marijuana X High				-1.560***	-1.558***	-1.558***	-1.608***	-1.669***	-1.671***
				(0.377)	(0.379)	(0.379)	(0.397)	(0.396)	(0.396)
Participant Characteristics (Ref. = Non-White, Female)									
White					0.0384	0.0702	0.280	0.337	0.341
					(0.236)	(0.261)	(0.245)	(0.258)	(0.258)
Male					0.275	0.411	0.469	0.491	0.475
					(0.214)	(0.574)	(0.532)	(0.620)	(0.617)
White X Male						-0.179	-0.177	-0.0941	-0.0926
						(0.617)	(0.591)	(0.667)	(0.663)
Rationale Factors									
Authoritarianism							0.886***	1.244***	1.256***
							(0.0918)	(0.273)	(0.279)

							(0.0963)	(0.349)	(0.345)
Authoritarianism X Protectiveness								-0.261	-0.262
								(0.366)	(0.364)
Participant Characteristics X Rationale Factors									
White X Authoritarianism								-0.349	-0.361
								(0.287)	(0.292)
Male X Authoritarianism								-0.436	-0.433
								(0.601)	(0.601)
White X Male X Authoritarianism								-0.121	-0.124
								(0.706)	(0.705)
White X Protectiveness								-0.199	-0.199
								(0.359)	(0.355)
Male X Protectiveness								-0.108	-0.101
								(0.612)	(0.609)
White X Male X Protectiveness								0.568	0.558
								(0.650)	(0.647)
White X Authoritarianism X Protectiveness								0.296	0.300
								(0.381)	(0.379)
Male X Authoritarianism X Protectiveness								0.446	0.479
								(0.949)	(0.946)
White X Male X Authoritarianism X Protectiveness								-0.448	-0.479
								(0.980)	(0.976)
Observations	849	849	849	849	849	849	849	849	849

Source: Educator survey
*** p<0.01, ** p<0.05, * p<0.1

Figure 1. Coefficient plot of ordered logistic regression predicting disciplinary response, clustered by participant, using educator sample

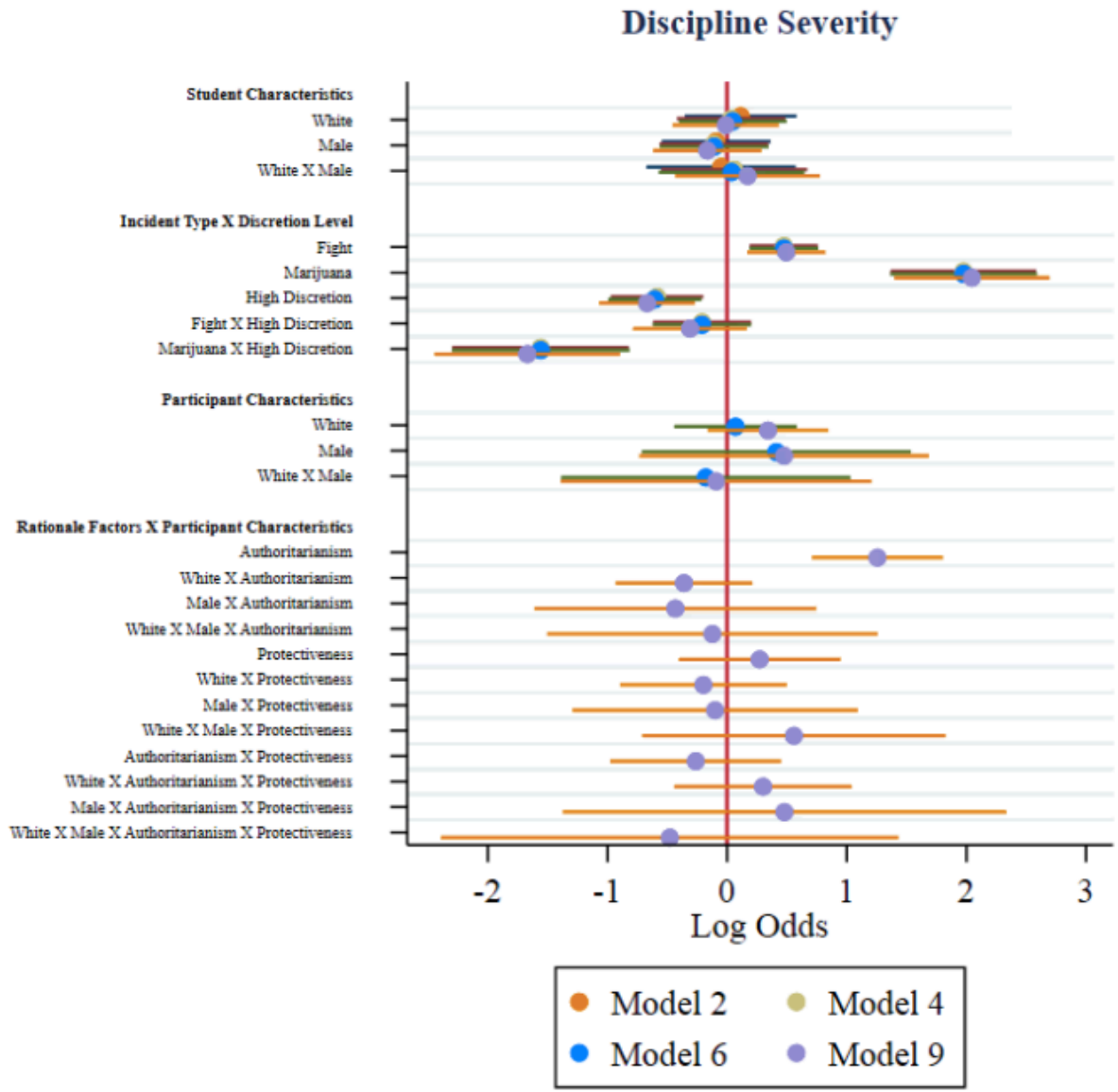


Table 7: Logistic regression predicting punishment, clustered by participant, using educator sample

VARIABLES	Model 1 B	Model 2 B	Model 3 B	Model 4 B	Model 5 B	Model 6 B	Model 7 B	Model 8 B	Model 9 B
Student Characteristics (Ref. = Black, Girl)									
White	-0.182	-0.110	-0.156	-0.156	-0.120	-0.122	-0.119	-0.149	-0.134
	(0.302)	(0.447)	(0.455)	(0.456)	(0.454)	(0.454)	(0.490)	(0.343)	(0.498)
Male	-0.0488	0.0265	0.0246	0.0247	0.0293	0.0331	-0.137	-0.111	-0.0944
	(0.299)	(0.464)	(0.474)	(0.474)	(0.473)	(0.473)	(0.517)	(0.343)	(0.541)
White X Male		-0.134	-0.101	-0.101	-0.173	-0.186	-0.0263		-0.0291
		(0.606)	(0.622)	(0.624)	(0.623)	(0.621)	(0.667)		(0.696)
Incident Type (Ref. = Defiance)									
Fight			0.267	-0.583	-0.584	-0.584	-0.859	-0.884	-0.885
			(0.268)	(0.520)	(0.520)	(0.520)	(0.585)	(0.598)	(0.601)
Marijuana			-0.946***	-1.678***	-1.684***	-1.684***	-2.418***	-2.378***	-2.378***
			(0.271)	(0.524)	(0.525)	(0.525)	(0.598)	(0.612)	(0.614)
Discretion Level (Ref. = Low)									
High Discretion			-0.383	-1.285**	-1.327**	-1.330**	-1.502**	-1.491**	-1.491**
			(0.313)	(0.588)	(0.585)	(0.585)	(0.661)	(0.686)	(0.687)
Incident Type X Discretion Level									
Fight X High				1.328**	1.333**	1.333**	1.384*	1.514**	1.515**
				(0.611)	(0.612)	(0.612)	(0.716)	(0.727)	(0.729)
Marijuana X High				1.110*	1.110*	1.110*	1.181	1.184	1.184
				(0.619)	(0.621)	(0.621)	(0.750)	(0.781)	(0.782)
Participant Characteristics (Ref. = Non-White, Female)									
White					0.331	0.256	0.953**	0.949**	0.948**
					(0.364)	(0.392)	(0.451)	(0.460)	(0.464)
Male					1.015	0.528	0.407	0.863	0.866
					(0.630)	(0.786)	(0.678)	(1.169)	(1.167)
White X Male						0.795	0.763	11.38*	11.39*
						(1.281)	(1.297)	(6.671)	(6.680)

Rationale Factors									
Authoritarianism							1.452***	0.710**	0.709**
							(0.161)	(0.299)	(0.299)
Protectiveness							-0.194	0.133	0.132
							(0.161)	(0.414)	(0.413)
Authoritarianism X Protectiveness								-0.611	-0.610
								(0.537)	(0.535)
Participant Characteristics X Rationale Factors									
White X Authoritarianism								0.826**	0.827**
								(0.355)	(0.355)
Male X Authoritarianism								0.167	0.165
								(0.438)	(0.442)
White X Male X Authoritarianism								6.993*	6.996*
								(4.143)	(4.147)
White X Protectiveness								-0.439	-0.438
								(0.494)	(0.491)
Male X Protectiveness								0.0926	0.0928
								(0.437)	(0.436)
White X Male X Protectiveness								-4.268	-4.270
								(3.160)	(3.160)
White X Authoritarianism X Protectiveness								0.530	0.529
								(0.563)	(0.561)
Male X Authoritarianism X Protectiveness								-0.204	-0.210
								(0.820)	(0.824)
White X Male X Authoritarianism X Protectiveness								10.46*	10.47*
								(5.515)	(5.520)
Constant	2.474***	2.434***	3.003***	3.619***	3.283***	3.350***	4.008***	4.052***	4.045***
	(0.285)	(0.336)	(0.389)	(0.556)	(0.661)	(0.683)	(0.793)	(0.805)	(0.811)
Observations	849	849	849	849	849	849	849	849	849

Source: Educator sample

*** p<0.01, ** p<0.05, * p<0.1

Figure 2. Coefficient plot of logistic regression predicting punishment, clustered by participant, using educator sample

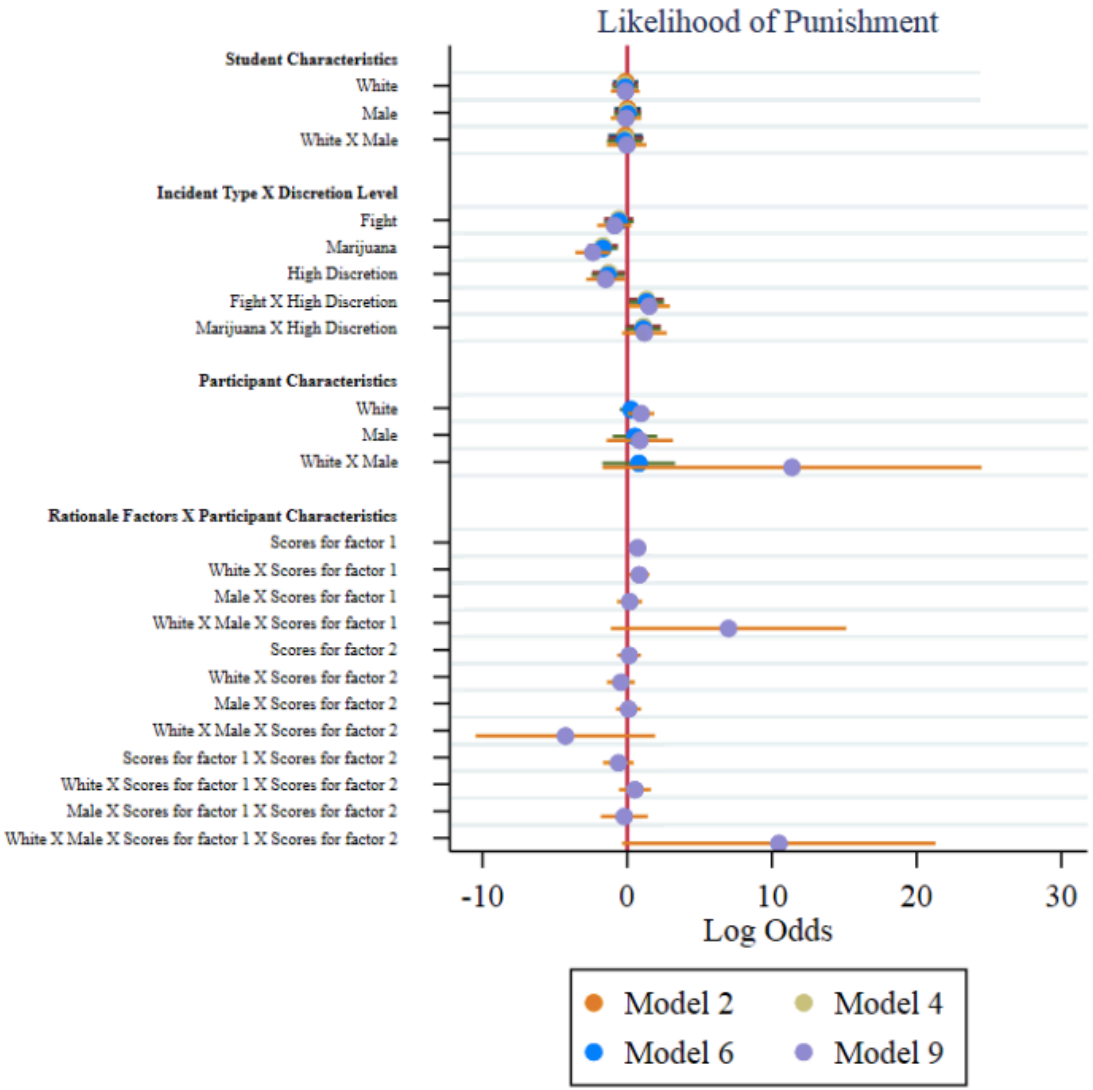


Table 8: Logistic regression predicting exclusionary discipline, clustered by participant, using educator sample

VARIABLES	Model 1 B	Model 2 B	Model 3 B	Model 4 B	Model 5 B	Model 6 B	Model 7 B	Model 8 B	Model 9 B
Student Characteristics (Ref. = Black, Girl)									
White	0.139	0.121	0.00570	0.00572	0.0170	0.0181	0.0487	0.145	0.0154
	(0.174)	(0.251)	(0.257)	(0.258)	(0.257)	(0.257)	(0.268)	(0.184)	(0.263)
Male	-0.0505	-0.0680	-0.0814	-0.0815	-0.0799	-0.0822	-0.0697	-0.00155	-0.126
	(0.174)	(0.254)	(0.260)	(0.260)	(0.262)	(0.260)	(0.272)	(0.184)	(0.269)
White X Male		0.0335	0.144	0.144	0.121	0.129	0.126		0.246
		(0.348)	(0.356)	(0.356)	(0.356)	(0.357)	(0.363)		(0.364)
Incident Type (Ref. = Defiance)									
Fight			0.625***	0.896***	0.897***	0.898***	0.985***	1.021***	1.023***
			(0.161)	(0.253)	(0.253)	(0.254)	(0.283)	(0.291)	(0.291)
Marijuana			0.560***	0.646***	0.647***	0.647***	0.738***	0.774***	0.772***
			(0.160)	(0.236)	(0.236)	(0.236)	(0.272)	(0.275)	(0.276)
Discretion Level (Ref. = Low)									
High Discretion			-1.080***	-0.877***	-0.890***	-0.891***	-1.023***	-1.056***	-1.059***
			(0.181)	(0.245)	(0.246)	(0.246)	(0.265)	(0.269)	(0.269)
Incident Type X Discretion Level									
Fight X High				-0.477	-0.477	-0.477	-0.560	-0.616*	-0.620*
				(0.329)	(0.329)	(0.329)	(0.365)	(0.374)	(0.375)
Marijuana X High				-0.171	-0.171	-0.171	-0.208	-0.245	-0.247
				(0.316)	(0.316)	(0.316)	(0.339)	(0.344)	(0.345)
Participant Characteristics (Ref. = Non-White, Female)									
White					0.0826	0.143	0.390	0.537	0.547
					(0.265)	(0.288)	(0.277)	(0.332)	(0.333)
Male					0.233	0.488	0.591	0.469	0.442
					(0.286)	(0.669)	(0.706)	(0.764)	(0.764)
White X Male						-0.343	-0.365	-0.0696	-0.0660
						(0.740)	(0.785)	(0.829)	(0.826)
Rationale Factors									

Authoritarianism							0.765***	1.459***	1.471***
							(0.107)	(0.382)	(0.389)
Protectiveness							0.175*	0.128	0.128
							(0.102)	(0.295)	(0.296)
Authoritarianism X Protectiveness								-0.179	-0.190
								(0.367)	(0.373)
Participant Characteristics X Rationale Factors									
White X Authoritarianism								-0.716*	-0.729*
								(0.396)	(0.403)
Male X Authoritarianism								-0.346	-0.338
								(0.665)	(0.670)
White X Male X Authoritarianism								-0.0358	-0.0356
								(0.805)	(0.808)
White X Protectiveness								0.00299	-0.000883
								(0.304)	(0.306)
Male X Protectiveness								-0.00614	-0.00223
								(0.679)	(0.672)
White X Male X Protectiveness								0.696	0.686
								(0.717)	(0.710)
White X Authoritarianism X Protectiveness								0.269	0.285
								(0.380)	(0.386)
Male X Authoritarianism X Protectiveness								1.013	1.072
								(0.750)	(0.755)
White X Male X Authoritarianism X Protectiveness								-1.046	-1.094
								(0.793)	(0.794)
Constant	0.338**	0.348*	0.577**	0.472*	0.383	0.329	0.161	-0.0242	0.0384
	(0.155)	(0.182)	(0.235)	(0.250)	(0.360)	(0.377)	(0.374)	(0.400)	(0.417)
Observations	849	849	849	849	849	849	849	849	849

Source: Educator survey
*** p<0.01, ** p<0.05, * p<0.1

Figure 3. Coefficient plot of logistic regression predicting exclusionary discipline, clustered by participant

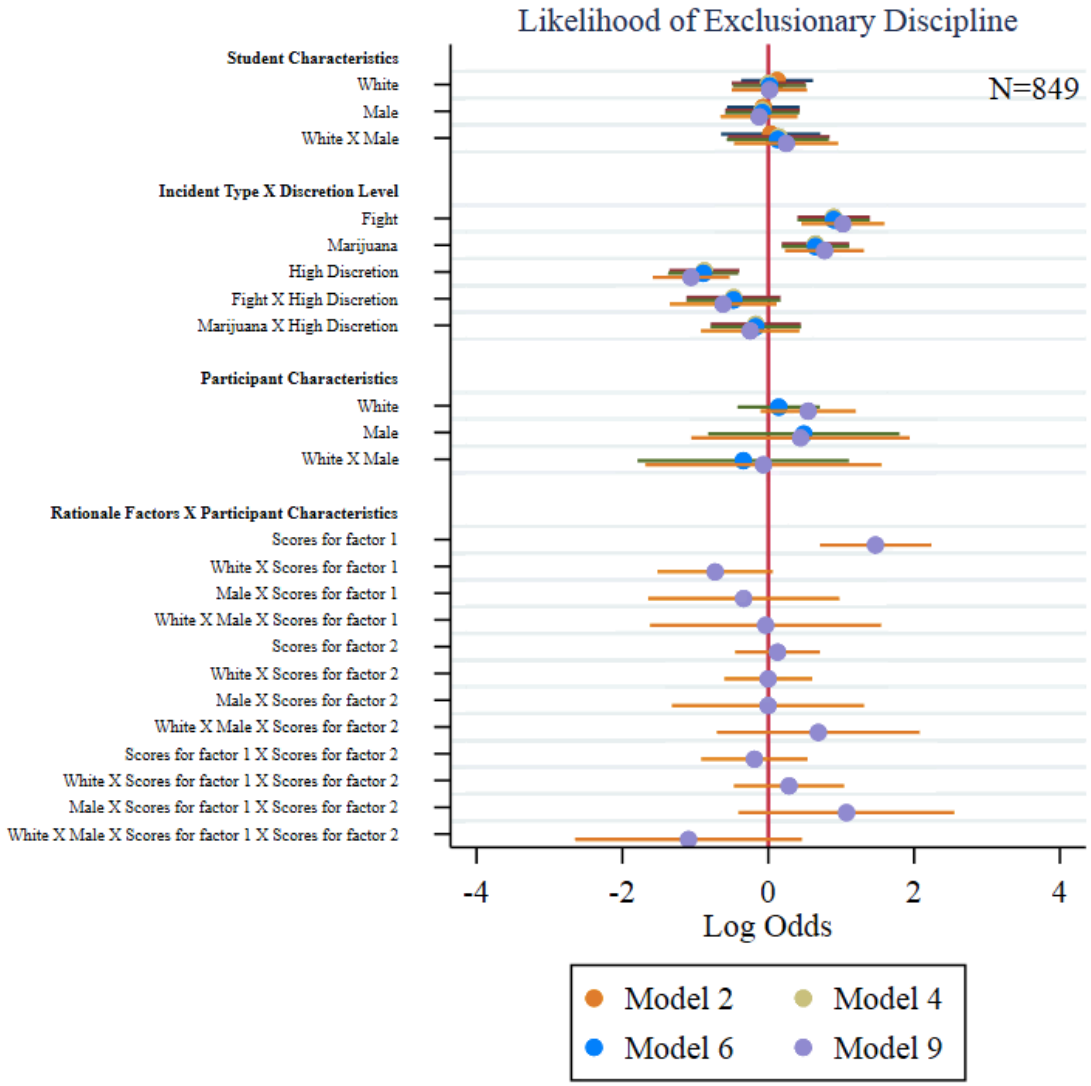


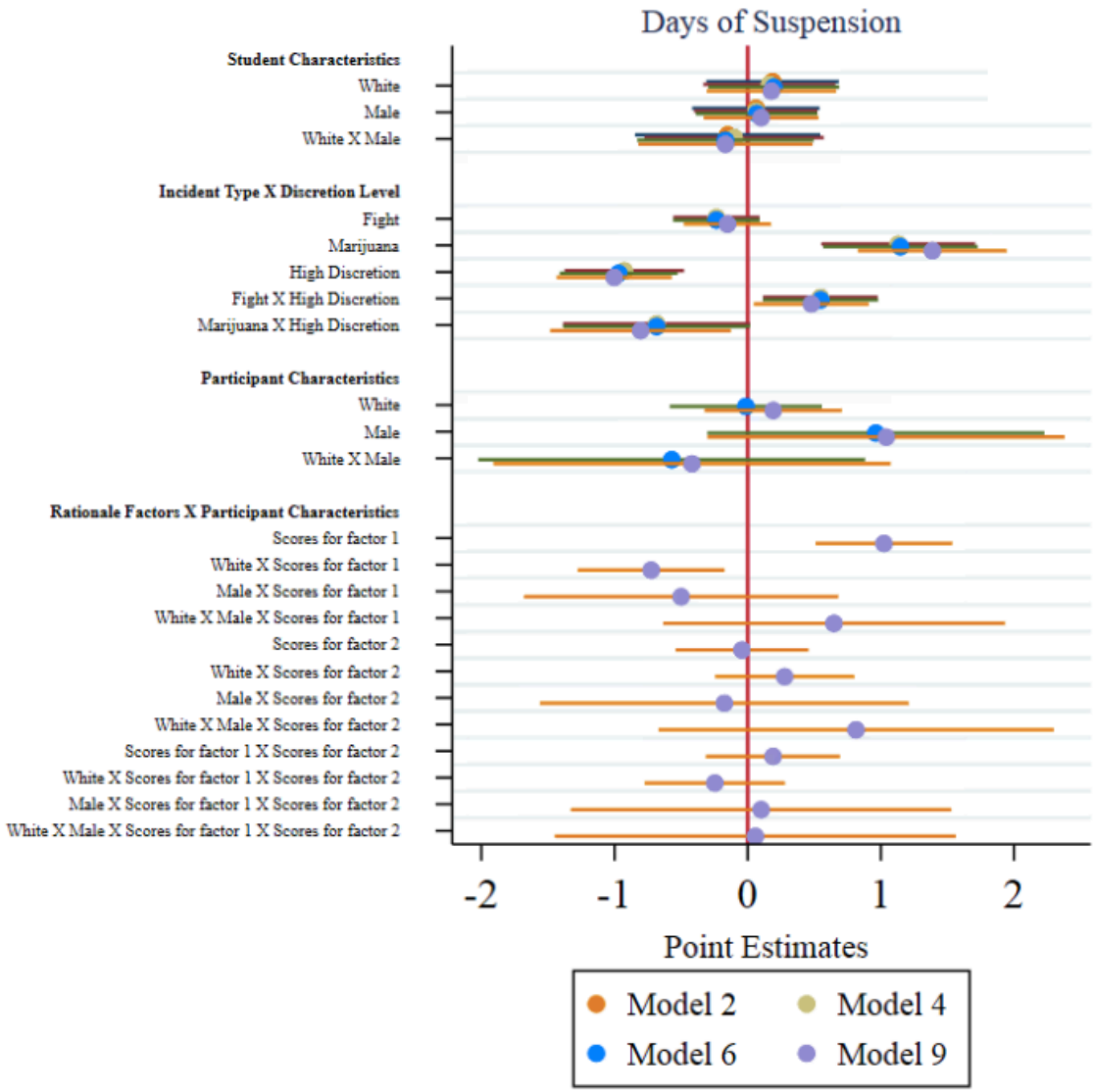
Table 9: Regression of days of suspension, clustered by participant, using educator sample

VARIABLES	Model 1 B	Model 2 B	Model 3 B	Model 4 B	Model 5 B	Model 6 B	Model 7 B	Model 8 B	Model 9 B
Student Characteristics (Ref. = Black, Girl)									
White	0.106	0.187	0.128	0.162	0.190	0.196	0.160	0.0872	0.177
	(0.177)	(0.252)	(0.252)	(0.252)	(0.252)	(0.249)	(0.252)	(0.167)	(0.247)
Male	-0.0170	0.0617	0.0532	0.0609	0.0670	0.0676	0.0877	0.0127	0.100
	(0.177)	(0.243)	(0.235)	(0.235)	(0.235)	(0.232)	(0.232)	(0.164)	(0.219)
White X Male		-0.150	-0.0740	-0.102	-0.173	-0.168	-0.180		-0.168
		(0.353)	(0.342)	(0.342)	(0.338)	(0.336)	(0.332)		(0.333)
Incident Type (Ref. = Defiance)									
Fight			0.0296	-0.235	-0.236	-0.235	-0.161	-0.153	-0.153
			(0.112)	(0.164)	(0.165)	(0.165)	(0.162)	(0.166)	(0.166)
Marijuana			0.652***	1.131***	1.140***	1.147***	1.364***	1.383***	1.387***
			(0.173)	(0.293)	(0.295)	(0.295)	(0.284)	(0.285)	(0.284)
Discretion Level (Ref. = Low)									
High Discretion			-0.842***	-0.927***	-0.968***	-0.969***	-0.983***	-1.005***	-1.002***
			(0.175)	(0.228)	(0.224)	(0.225)	(0.218)	(0.219)	(0.220)
Incident Type X Discretion Level									
Fight X High				0.546**	0.550**	0.547**	0.485**	0.476**	0.478**
				(0.219)	(0.219)	(0.220)	(0.211)	(0.219)	(0.219)
Marijuana X High				-0.684*	-0.683*	-0.684*	-0.758**	-0.798**	-0.805**
				(0.357)	(0.357)	(0.357)	(0.343)	(0.346)	(0.345)
Participant Characteristics (Ref. = Non-White, Female)									
White					-0.122	-0.0141	0.0632	0.195	0.191
					(0.270)	(0.290)	(0.276)	(0.264)	(0.262)
Male					0.527*	0.962	1.061	1.017	1.039
					(0.314)	(0.643)	(0.651)	(0.684)	(0.681)
White X Male						-0.570	-0.575	-0.413	-0.419
						(0.738)	(0.736)	(0.759)	(0.758)
Rationale Factors									
Authoritarianism							0.364***	1.033***	1.023***
							(0.0855)	(0.263)	(0.261)
Protectiveness							0.251***	-0.0445	-0.0432
							(0.0852)	(0.257)	(0.254)
Authoritarianism X Protectiveness								0.184	0.190
								(0.255)	(0.255)

Participant Characteristics X Rationale Factors									
White X Authoritarianism								-0.737***	-0.725***
								(0.283)	(0.280)
Male X Authoritarianism								-0.507	-0.500
								(0.607)	(0.600)
White X Male X Authoritarianism								0.659	0.648
								(0.660)	(0.652)
White X Protectiveness								0.274	0.277
								(0.266)	(0.265)
Male X Protectiveness								-0.162	-0.176
								(0.706)	(0.703)
White X Male X Protectiveness								0.799	0.814
								(0.756)	(0.753)
White X Authoritarianism X Protectiveness								-0.235	-0.247
								(0.266)	(0.267)
Male X Authoritarianism X Protectiveness								0.148	0.100
								(0.733)	(0.726)
White X Male X Authoritarianism X Protectiveness								0.00696	0.0568
								(0.768)	(0.764)
Constant	3.278***	3.235***	3.528***	3.554***	3.614***	3.515***	3.360***	3.274***	3.226***
	(0.155)	(0.179)	(0.207)	(0.220)	(0.363)	(0.369)	(0.353)	(0.309)	(0.321)
Observations	629	629	629	629	629	629	629	629	629
R-squared	0.001	0.001	0.074	0.091	0.103	0.105	0.159	0.183	0.184

Source: Educator sample
Note: Regression analysis includes only cases in which the participant recommend in-school or out-of-school suspension, resulting in 629 observations.
*** p<0.01, ** p<0.05, * p<0.1

Figure 4. Coefficient plot of regression predicting days of suspension, clustered by participant, using educator sample



APPENDIX B: TABLES AND FIGURES FOR MTURK SAMPLE

Table 10. Participant characteristics in MTurk sample

	Number	Percentage
Race*		
White or Caucasian	187	78.2%
Black or African American	24	10.0%
Hispanic or Latino	15	6.3%
Asian	19	7.9%
Other	5	2.1%
Gender		
Woman	138	57.7%
Man	101	42.3%
Age		
Younger than 30	92	38.5%
30-39	94	39.3%
40-49	28	11.7%
50-59	19	7.9%
60-69	6	2.5%
70 and older	0	0.0%

Observations (N)=239

*Racial categories were not exclusive, so cumulative tallies will be greater than N.

Table 11. Distribution of independent variables in educator sample

	Student Race and Gender					
	Black or African American			**White or Caucasian**		
Discipline policy	Girl	Boy	Total	Girl	Boy	Total
Low discretion	87	96	183	93	102	195
High discretion	93	87	180	84	75	159
Totals	180	183	363	177	177	354

Observations (N) = 717

Table 12. Distribution of participants' recommended disciplinary actions

	Frequency	Percent
None	15	2.1%
Detention	60	8.4%
In-School Suspension	166	23.2%
1-2 days	48	28.9%
3-4 days	43	25.9%
5-7 days	75	45.2%
Out-of-School Suspension	347	48.4%
1-2 days	21	6.1%
3-4 days	150	43.2%
5-7 days	176	50.7%
Expulsion	129	18.0%
N	717	100.0%

Table 13. Educators' average ratings, on a scale of 1-5, of the importance of six rationales

Rationale	Average rating
Providing sufficient punishment for the student's behavior	4.06
Preventing future misbehavior from the student	4.06
Discouraging other students from similar behavior	4.09
Minimizing disruption to the learning environment	3.90
Ensuring the safety of other students	3.87
Ensuring the safety of staff	3.66

Observations (N) = 717

Table 14. Factor loadings and unique variances for rationales in educator sample

Variable	Factor 1	Uniqueness
Providing a sufficient punishment for the student's misbehavior	**0.7333**	0.4622
Preventing future misbehavior from the student	**0.7963**	0.3660
Discouraging other students from similar behavior	**0.7742**	0.4006
Ensuring the safety of other students	**0.7574**	0.4263
Ensuring the safety of school staff	**0.7329**	0.4629
Minimizing disruption to the learning environment	**0.7813**	0.3895

* Those I consider influential, using a factor loading of 0.5 as the cut-off, are marked in bold font. It is clear that Factor 1 corresponds with the participants' rating of the importance of all six disciplinary factors. Thus, I refer to Factor 1 broadly as "Disciplinary Zeal."

Table 15. Ordered logistic regression predicting disciplinary response, clustered by participant, using MTurk sample

VARIABLES	Model 1 B	Model 2 B	Model 3 B	Model 4 B	Model 5 B	Model 6 B	Model 7 B	Model 8 B	Model 9 B
Student Characteristics (Ref. = Black, Female)									
White	0.151	0.0284	0.0313	0.0271	-0.00645	-0.0163	-0.0759	0.0892	-0.0743
	(0.160)	(0.216)	(0.212)	(0.217)	(0.216)	(0.216)	(0.224)	(0.171)	(0.221)
Male	0.0389	-0.0854	-0.106	-0.116	-0.156	-0.162	-0.237	-0.0434	-0.214
	(0.160)	(0.237)	(0.248)	(0.252)	(0.253)	(0.254)	(0.262)	(0.171)	(0.264)
White X Male		0.249	0.224	0.230	0.330	0.335	0.375		0.339
		(0.320)	(0.323)	(0.330)	(0.334)	(0.334)	(0.342)		(0.347)
Incident Type (Ref. = Defiance)									
Fight			0.337***	0.609***	0.599***	0.597***	0.610***	0.607***	0.610***
			(0.126)	(0.143)	(0.145)	(0.145)	(0.146)	(0.145)	(0.145)
Marijuana			1.193***	2.241***	2.238***	2.237***	2.548***	2.541***	2.542***
			(0.205)	(0.278)	(0.284)	(0.284)	(0.305)	(0.306)	(0.306)
Discretion Level (Ref. = Low Discretion)									
High Discretion			-0.727***	0.0455	0.0754	0.0734	-0.0754	-0.0698	-0.0675
			(0.173)	(0.224)	(0.226)	(0.226)	(0.226)	(0.228)	(0.228)
Incident Type X Discretion Level									
Fight X High Discretion				-0.554**	-0.573**	-0.576**	-0.534**	-0.521*	-0.529**
				(0.266)	(0.269)	(0.269)	(0.269)	(0.268)	(0.269)
Marijuana X High Discretion				-2.096***	-2.123***	-2.120***	-2.164***	-2.145***	-2.151***
				(0.376)	(0.381)	(0.380)	(0.387)	(0.389)	(0.390)
Participant Characteristics (Ref. = Non-White, Female)									
White					0.312	0.156	0.419	0.594	0.601
					(0.219)	(0.415)	(0.410)	(0.503)	(0.520)
Male					-0.338**	-0.520	-0.266	-0.0600	-0.0823
					(0.170)	(0.452)	(0.444)	(0.531)	(0.548)
White X Male						0.218	0.0351	-0.168	-0.147
						(0.485)	(0.480)	(0.563)	(0.579)
Rationale Factor									
Disciplinary Zeal							0.594***	0.980**	0.967**
							(0.119)	(0.443)	(0.457)
Rationale Factor X Participant Characteristics									
White X Disciplinary Zeal								-0.474	-0.454
								(0.473)	(0.490)
Male X Disciplinary Zeal								-0.525	-0.502
								(0.556)	(0.564)

White X Male X Disciplinary Zeal								0.726	0.691
								(0.608)	(0.618)
Observations	717	717	717	717	717	717	717	717	717

Source: MTurk sample
*** p<0.01, ** p<0.05, * p<0.1

Figure 5. Coefficient Plot of Ordered logistic regression predicting disciplinary response, clustered by participant, using MTurk sample

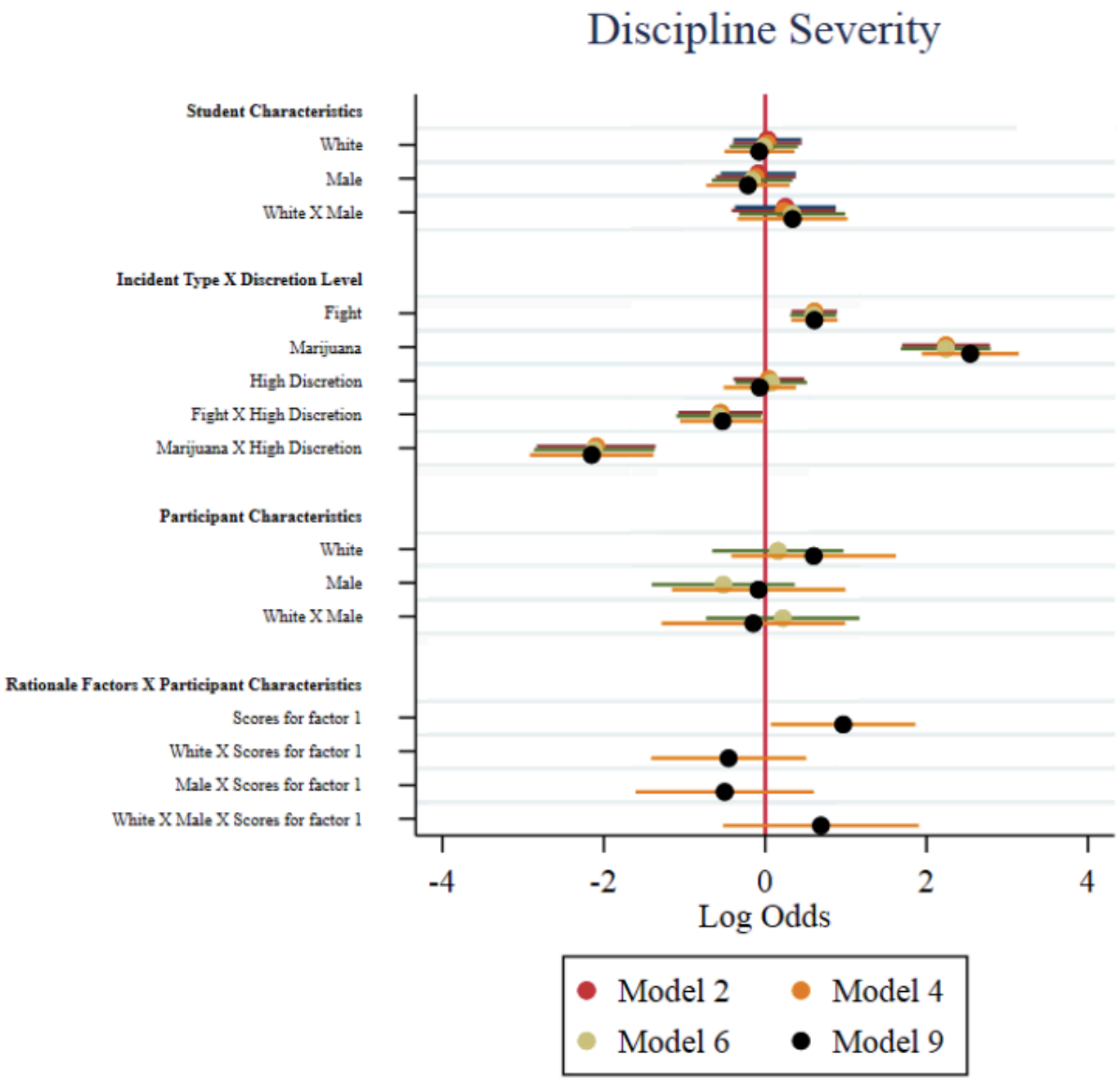


Table 16: Logistic regression predicting punishment, clustered by participant, using MTurk sample

VARIABLES	Model 1 B	Model 2 B	Model 3 B	Model 4 B	Model 5 B	Model 6 B	Model 7 B	Model 8 B	Model 9 B
Student Characteristics (Ref. = Black, Female)									
White	1.388**	-0.0170	-0.0506	-0.0499	0.00558	0.00558	-0.706	0.928	-1.375
	(0.692)	(0.996)	(0.996)	(0.995)	(1.004)	(1.004)	(0.788)	(0.777)	(0.844)
Male	-1.025	-1.638**	-1.706**	-1.702**	-1.650**	-1.650**	-2.879***	-1.776**	-3.258***
	(0.642)	(0.825)	(0.824)	(0.822)	(0.817)	(0.817)	(0.797)	(0.783)	(0.702)
White X Male		2.337	2.367	2.364	2.160	2.160	2.670*		3.404**
		(1.475)	(1.483)	(1.481)	(1.481)	(1.481)	(1.608)		(1.669)
Incident Type (Ref. = Defiance)									
Fight			-0.415	0.265	0.258	0.258	-0.0529	-0.0432	-0.221
			(0.725)	(1.370)	(1.384)	(1.384)	(1.578)	(1.487)	(1.529)
Marijuana			-1.687**	-1.170	-1.191	-1.191	-0.117	-0.0297	-0.122
			(0.720)	(0.743)	(0.753)	(0.753)	(0.758)	(0.770)	(0.779)
Discretion Level (Ref. = Low Discretion)									
High Discretion			-0.834	-0.540	-0.508	-0.508	-1.844**	-1.428*	-1.718*
			(0.624)	(0.671)	(0.688)	(0.688)	(0.910)	(0.789)	(0.881)
Incident Type X Discretion Level									
Fight X High Discretion				-0.265	-0.258	-0.258	0.0234	0.0144	0.105
				(1.296)	(1.308)	(1.308)	(1.466)	(1.385)	(1.454)
Marijuana X High Discretion				0	0	0	0	0	0
				(0)	(0)	(0)	(0)	(0)	(0)
Participant Characteristics (Ref. = Non-White, Female)									
White					-	-	-	-	-
Male					-0.548	-0.548	-0.515	0.698	0.948
					(0.689)	(0.689)	(0.803)	(1.253)	(1.303)
White X Male						0	0	0	0
						(0)	(0)	(0)	(0)
Rationale Factor									

Disciplinary Zeal							1.889***	1.392**	1.328**
							(0.433)	(0.668)	(0.624)
Participant Characteristics X Rationale Factor									
White X Disciplinary Zeal								0	0
								(0)	(0)
Male X Disciplinary Zeal								0.788	1.066
								(0.691)	(0.680)
White X Male X Disciplinary Zeal								0	0
								(0)	(0)
Constant	4.011***	4.489***	5.965***	5.263***	5.331***	5.331***	8.387***	6.401***	7.962***
	(0.471)	(0.705)	(0.761)	(0.873)	(0.888)	(0.888)	(1.976)	(1.939)	(1.855)
Observations	717	717	717	591	463	463	463	463	463

*** p<0.01, ** p<0.05, * p<0.1

Notes: Defiance X Low Discretion perfectly predicted punishment; thus, 126 observations were excluded from the analysis.
Similarly, for non-white participants, this racial characteristic perfectly predicted punishment, so 128 observations were excluded from the analysis.
Meanwhile, participant whiteness resulted in collinearity, so these data points above are marked as "-" and result in values of zero in their interactions.
Source: MTurk sample

Figure 6. Coefficient plot of logistic regression predicting likelihood of punishment, clustered by participant, using MTurk sample

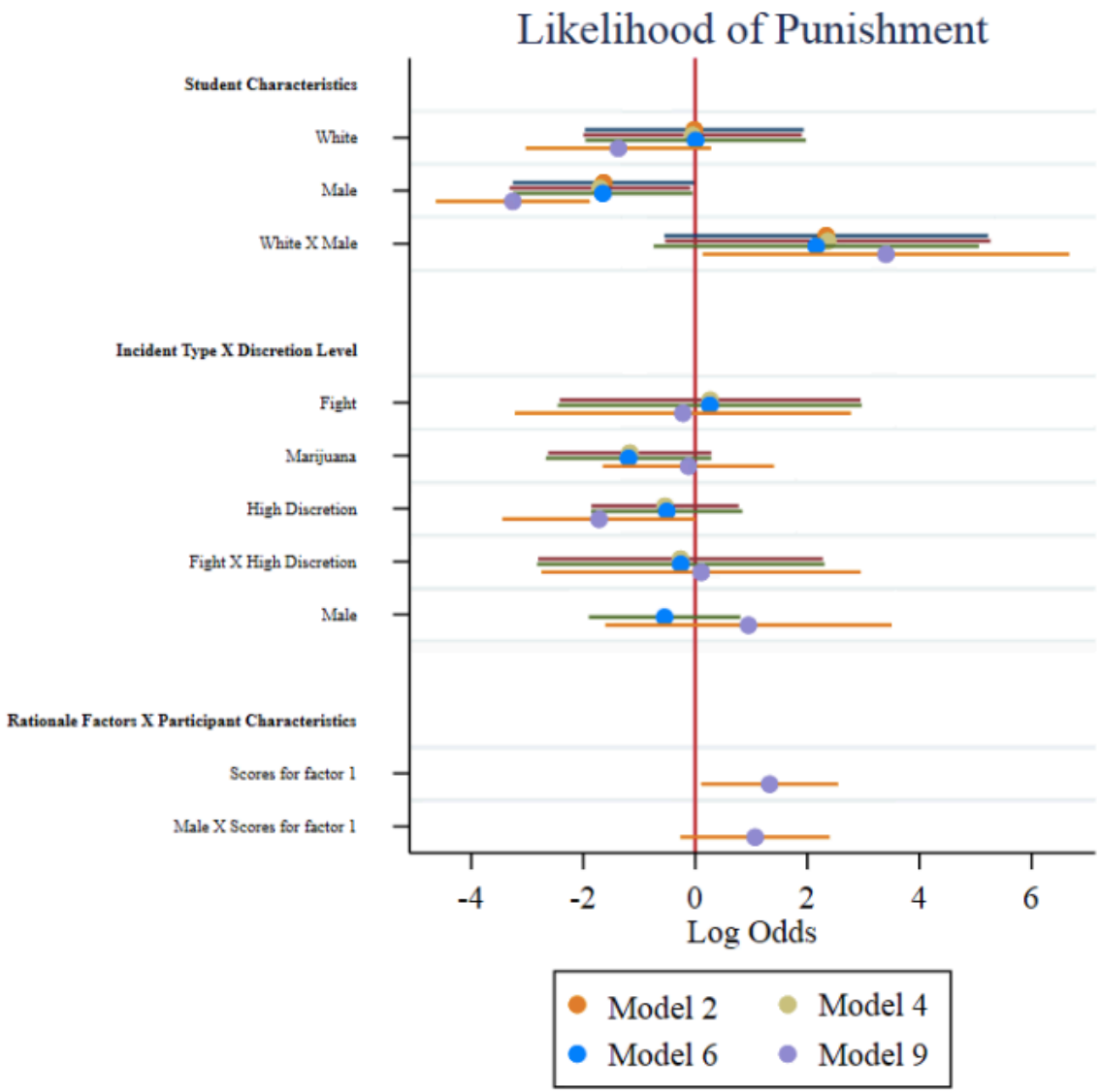


Table 17: Logistic regression predicting exclusionary discipline, clustered by participant, using MTurk sample

VARIABLES	Model 1 B	Model 2 B	Model 3 B	Model 4 B	Model 5 B	Model 6 B	Model 7 B	Model 8 B	Model 9 B
Student Characteristics (Ref. = Black, Girl)									
White	0.150	-0.152	-0.192	-0.193	-0.238	-0.259	-0.296	0.0689	-0.313
	(0.190)	(0.267)	(0.267)	(0.269)	(0.269)	(0.270)	(0.284)	(0.207)	(0.280)
Male	-0.0744	-0.362	-0.412	-0.416	-0.469	-0.481*	-0.521*	-0.127	-0.494*
	(0.190)	(0.270)	(0.280)	(0.284)	(0.287)	(0.289)	(0.300)	(0.203)	(0.300)
White X Male		0.596	0.618	0.626	0.767*	0.776**	0.771*		0.765*
		(0.377)	(0.384)	(0.388)	(0.392)	(0.393)	(0.407)		(0.406)
Incident Type (Ref. = Defiance)									
Fight			0.634***	1.208***	1.234***	1.236***	1.331***	1.332***	1.338***
			(0.168)	(0.245)	(0.251)	(0.252)	(0.263)	(0.260)	(0.260)
Marijuana			0.550***	1.101***	1.125***	1.127***	1.475***	1.453***	1.458***
			(0.182)	(0.240)	(0.245)	(0.246)	(0.273)	(0.275)	(0.277)
Discretion Level (Ref. = Low Discretion)									
High Discretion			-0.751***	-0.0842	-0.0609	-0.0611	-0.224	-0.234	-0.228
			(0.195)	(0.265)	(0.270)	(0.270)	(0.279)	(0.280)	(0.282)
Incident Type X Discretion Level									
Fight X High Discretion				-1.098***	-1.121***	-1.123***	-1.139***	-1.128***	-1.133***
				(0.336)	(0.343)	(0.344)	(0.354)	(0.353)	(0.355)
Marijuana X High Discretion				-1.065***	-1.088***	-1.089***	-1.162***	-1.130***	-1.137***
				(0.358)	(0.367)	(0.367)	(0.375)	(0.379)	(0.383)
Participant Characteristics (Ref. = Non-White, Female)									
White					0.586**	0.272	0.511	0.807	0.853
					(0.235)	(0.522)	(0.507)	(0.639)	(0.655)
Male					-0.354*	-0.704	-0.496	-0.125	-0.141
					(0.210)	(0.539)	(0.522)	(0.648)	(0.667)
White X Male						0.428	0.331	-0.0225	-0.00582
						(0.582)	(0.572)	(0.689)	(0.706)
Rationale Factor									
Discretionary Zeal							0.568***	1.684***	1.700***
							(0.118)	(0.575)	(0.577)

Rationale Factor X Participant Characteristics									
White X Discretionary Zeal								-1.202**	-1.220**
								(0.602)	(0.606)
Male X Discretionary Zeal								-1.259*	-1.254*
								(0.644)	(0.645)
White X Male X Discretionary Zeal								1.407**	1.391**
								(0.684)	(0.686)
Constant	0.645***	0.795***	0.829***	0.504**	0.260	0.543	0.311	-0.192	-0.0433
	(0.156)	(0.183)	(0.255)	(0.257)	(0.366)	(0.566)	(0.545)	(0.637)	(0.670)
Observations	717	717	717	717	717	717	717	717	717

Source: MTurk sample.
*** p<0.01, ** p<0.05, * p<0.1

Figure 7. Coefficient plot of logistic regression predicting exclusionary discipline, clustered by participant

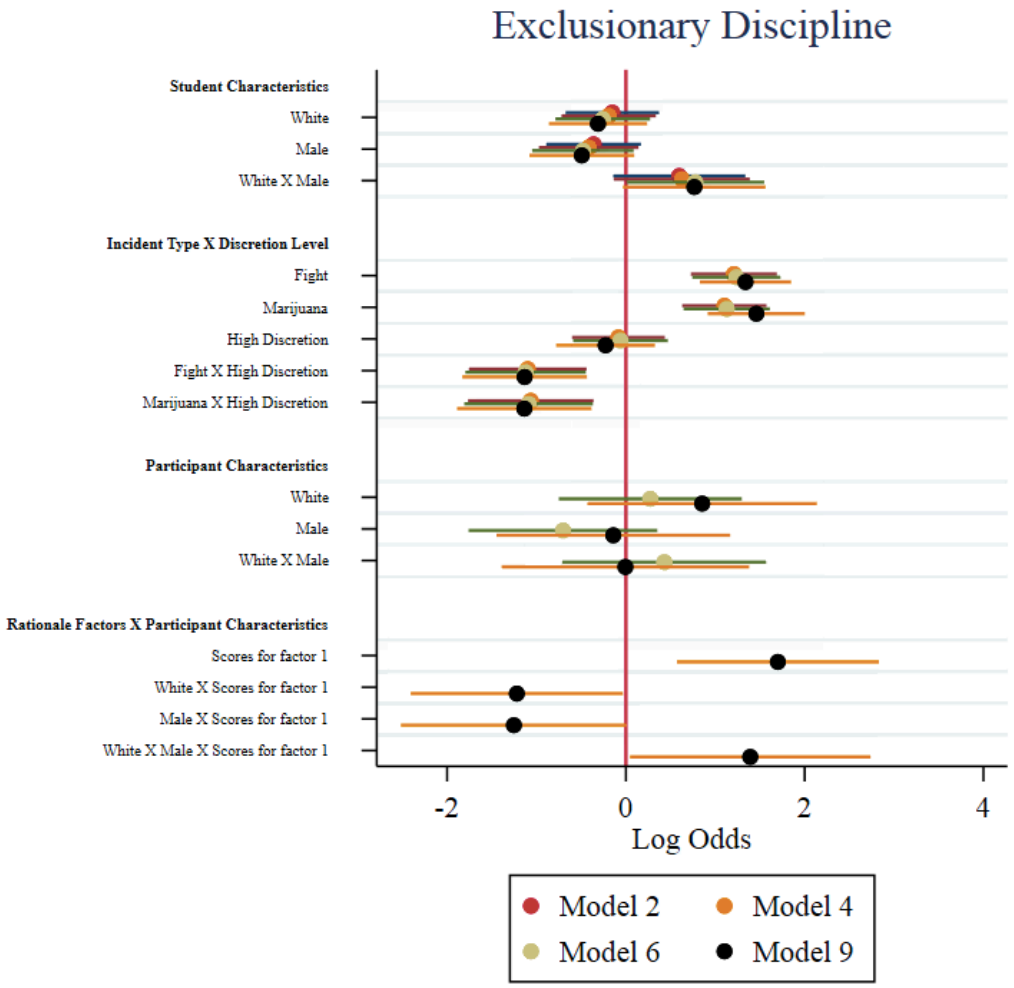


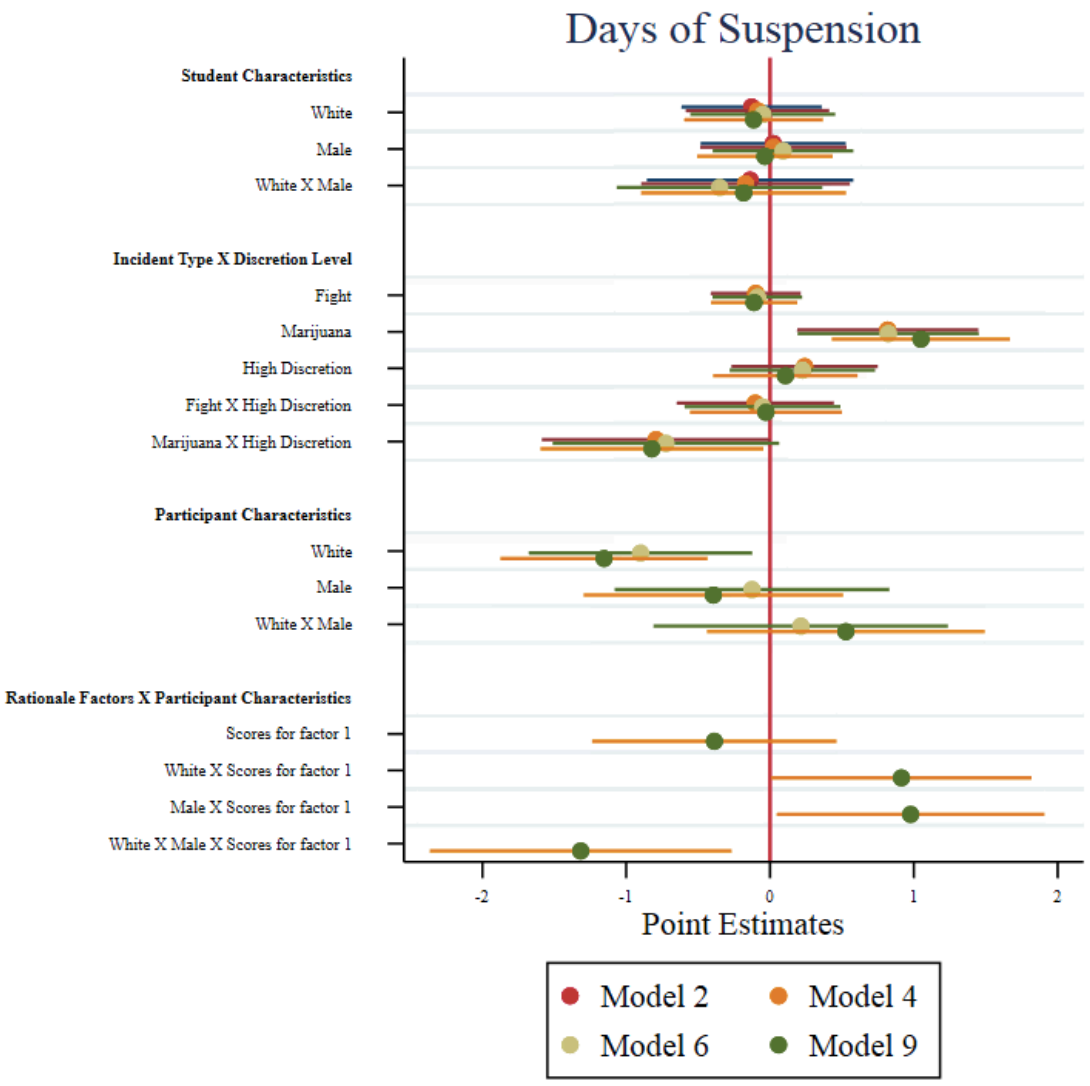
Table 18: Regression predicting days of suspension, clustered by participant, using MTurk sample

VARIABLES	Model 1 B	Model 2 B	Model 3 B	Model 4 B	Model 5 B	Model 6 B	Model 7 B	Model 8 B	Model 9 B
Student Characteristics (Ref. = Black, Female)									
White	-0.196	-0.128	-0.101	-0.0873	-0.0418	-0.0506	-0.0870	-0.202	-0.114
	(0.182)	(0.248)	(0.250)	(0.253)	(0.252)	(0.256)	(0.246)	(0.178)	(0.245)
Male	-0.0477	0.0217	0.0296	0.0235	0.0956	0.0903	0.0206	-0.128	-0.0360
	(0.182)	(0.256)	(0.258)	(0.258)	(0.248)	(0.248)	(0.245)	(0.175)	(0.239)
White X Male		-0.139	-0.159	-0.170	-0.354	-0.351	-0.303		-0.184
		(0.365)	(0.367)	(0.368)	(0.363)	(0.363)	(0.360)		(0.362)
Incident Type (Ref. = Defiance)									
Fight			-0.149	-0.0993	-0.0857	-0.0881	-0.0988	-0.113	-0.112
			(0.132)	(0.159)	(0.158)	(0.158)	(0.152)	(0.152)	(0.153)
Marijuana			0.368*	0.818**	0.824**	0.822**	1.028***	1.044***	1.049***
			(0.199)	(0.320)	(0.319)	(0.320)	(0.314)	(0.315)	(0.314)
Discretion Level (Ref. = Low Discretion)									
High Discretion			0.0211	0.241	0.228	0.225	0.127	0.106	0.107
			(0.191)	(0.258)	(0.257)	(0.257)	(0.253)	(0.254)	(0.255)
Incident Type X Discretion Level									
Fight X High Discretion				-0.103	-0.0488	-0.0524	-0.0202	-0.0354	-0.0288
				(0.277)	(0.274)	(0.275)	(0.266)	(0.267)	(0.268)
Marijuana X High Discretion				-0.793*	-0.728*	-0.724*	-0.767*	-0.823**	-0.821**
				(0.403)	(0.400)	(0.400)	(0.394)	(0.393)	(0.393)
Participant Characteristics (Ref. = Non-White, Female)									
White					-0.751***	-0.901**	-0.704*	-1.154***	-1.155***
					(0.254)	(0.394)	(0.419)	(0.373)	(0.366)
Male					0.0553	-0.126	0.0758	-0.414	-0.394
					(0.180)	(0.485)	(0.505)	(0.461)	(0.458)
White X Male						0.214	0.0488	0.547	0.527
						(0.519)	(0.533)	(0.495)	(0.490)
Rationale Factor									
Discretionary Zeal							0.385***	-0.410	-0.388
							(0.0984)	(0.435)	(0.431)

Rationale Factor X Participant Characteristics									
White X Discretionary Zeal								0.941**	0.913**
								(0.459)	(0.459)
Male X Discretionary Zeal								1.010**	0.977**
								(0.475)	(0.473)
White X Male X Discretionary Zeal								-1.360**	-1.316**
								(0.526)	(0.532)
Constant	4.544***	4.511***	4.462***	4.363***	4.908***	5.050***	4.870***	5.395***	5.346***
	(0.155)	(0.178)	(0.205)	(0.220)	(0.322)	(0.429)	(0.443)	(0.404)	(0.406)
Observations	513	513	513	513	513	513	513	513	513
R-squared	0.004	0.004	0.018	0.027	0.058	0.058	0.099	0.113	0.114

Robust standard errors in parentheses
Source: MTurk sample
Note: Responses were dropped if the participant recommended neither in-school suspension nor out-of-school suspension, resulting in 513 observations for analysis.
*** p<0.01, ** p<0.05, * p<0.1

Figure 8. Coefficient plot of regression predicting days of suspension, clustered by participant, using MTurk sample



APPENDIX C: EDUCATOR SURVEY

Start of Block: Intro/Informed Consent

THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
COLLEGE OF ARTS AND SCIENCES
Department of Sociology

Dear educators:

This research project is designed to collect information on the way decisions are made about school discipline. The survey is being distributed to public school educators from across the United States, and the goal is to have 800 or more people participate. The survey should take you approximately 15 minutes to complete.

Your participation is voluntary and you may stop at any time without any penalty. You must be 18 or older and currently employed by a US public school or school district to participate. Your participation is completely anonymous. No identifying information will be collected, so participants' identities will not be known. Furthermore, individual responses will not be shared. Research results will be shared only in the aggregate (as a whole), such that no individual will be identifiable in any products of the research.

The Institutional Review Board at UNC-Chapel Hill has reviewed this study and determined it to be exempt from further consideration as it involves minimal possible risk to participants. If you have questions about the study, please contact me directly at sarahmd@live.unc.edu.

Thank you very much for your participation.

Kind regards,
Sarah M. Davis

End of Block: Intro/Informed Consent

Start of Block: Qualifying Questions

Are you 18 years of age or older?

- ○ Yes
- ○ No

Are you currently employed by a U.S. K-12 public school or public school district?

- ○ Yes
- ○ No

Survey Flow: If participant answers "no" to either question above, the survey automatically ends.

End of Block: Qualifying Questions

Start of Block: Directions

<u>Directions:</u>

In the survey that follows, you will be presented with a fictional story describing a student misbehavior incident. You will also be provided with a mock school discipline policy. Please use your best judgment using the information given to recommend a disciplinary action for the student.

Please read each section carefully. You will be asked questions about what you read.

End of Block: Directions

Survey Flow: Participants are given one of the two policies below which remains the same for the duration of their survey.

Start of Block: School Policy: High Discretion

School Policy

Please read carefully.

In this middle school, principals and teachers have full authority as provided by law to establish and enforce standards and rules as are necessary to create orderly schools and classrooms. According to the Student Code of Conduct, disciplinary consequences may include, but are not limited to, the following:

5. Detention
6. In-school suspension
7. Out-of-school suspension
8. Expulsion

This policy will remain visible as you continue with the survey.

End of Block: School Policy: High discretion

Start of Block: School Policy: Low discretion

School Policy

Please read carefully.

In this middle school, principals and teachers have full authority as provided by law to establish and enforce standards and rules as are necessary to create orderly schools and classrooms. According to the Student Code of Conduct, disciplinary consequences are assigned to students according to their offense, as outlined below:

10. Defiance or disrespect → detention
11. Inappropriate language → detention
12. Disruption → in-school suspension for one day
13. Leaving school without permission → in-school suspension for one day
14. Skipping → in-school suspension for one day
15. Communicating threats → out-of-school suspension for three days
16. Physical aggression toward another student → out-of-school suspension for four days
17. Assault of school employee → out-of-school suspension for five days
18. Possession of drugs: first offense → out-of-school suspension for seven days
19. Possession of drugs: second offense → expulsion

This policy will remain visible as you continue with the survey.

End of Block: School Policy: Low discretion

Survey Flow: The order of the three following incidents is randomized for each participant.

Start of Block: Incident 1

Imagine the following...

One day in [student name]'s 4th period class, [student name] refused to do [his/her] work and disrupted other students from doing theirs. The teacher asked [him/her] to sit at the back table so [he/she] could focus better, but [he/she] refused. When the teacher walked over to [student name]'s desk to talk with [him/her], [he/she] stood up and yelled, "I said I'm not going, you f***ing b****!" [He/She] then charged out of the classroom, shoving the teacher roughly out of the way, and left the building.

What disciplinary action would you recommend for [student name]?*

- ○ Detention
- ○ In-School Suspension
- ○ Out-of-School Suspension
- ○ Expulsion
- ○ None of the above

---

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

Survey Flow: If in-school or out-of-school suspension is selected….

How many days of suspension would you recommend?*

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7

---

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

Survey Flow: The order of the seven factors below is randomized and broken across two pages for ease of reading

How important were the following factors in making your disciplinary recommendation?

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Delivering a sufficient punishment for the student's misbehavior	○	○	○	○	○
Preventing future misbehavior from the student	○	○	○	○	○
Discouraging other students from similar behavior	○	○	○	○	○
Minimizing disruption to the learning environment	○	○	○	○	○
Ensuring the safety of other students	○	○	○	○	○
Ensuring the safety of school staff	○	○	○	○	○
Following the school discipline policy	○	○	○	○	○

End of Block: Incident 1

Start of Block: Incident 2

Now imagine that instead, this incident took place...

At lunch in the cafeteria, the School Resource Officer saw [student name] and another student stand up, get in each other's faces, and shout threats and insults at one another. Before the School Resource Officer could stop the fight, [student name] had pushed the other student to the ground and punched [him/her – him if the primary student is male, her if the primary student is female] in the face resulting in a bloody nose. After receiving statements from both students, it seems that the confrontation was over a shared romantic interest.

What disciplinary action would you recommend for [student name]?*

- ○ Detention
- ○ In-School Suspension
- ○ Out-of-School Suspension
- ○ Expulsion
- ○ None of the above

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

Survey Flow: If in-school or out-of-school suspension is selected….

How many days of suspension would you recommend?*

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7

---

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

How important were the following factors in making your disciplinary recommendation?

Survey Flow: The order of the seven factors below is randomized and broken across two pages for ease of reading

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Delivering a sufficient punishment for the student's misbehavior	○	○	○	○	○
Preventing future misbehavior from the student	○	○	○	○	○
Discouraging other students from similar behavior	○	○	○	○	○
Minimizing disruption to the learning environment	○	○	○	○	○
Ensuring the safety of other students	○	○	○	○	○
Ensuring the safety of school staff	○	○	○	○	○
Following the school discipline policy	○	○	○	○	○

End of Block: Incident 2

Start of Block: Incident 3

Finally, imagine that instead of either of the two previous scenarios, the following occurred...

As students boarded the school bus at the end of the school day, the driver noticed [student name] had bloodshot eyes and seemed unusually spacey. The driver called the front office and when they searched [student name]'s bag they found a small bag of marijuana and a pipe. [Student name] has been caught with marijuana once already this school year.

What disciplinary action would you recommend for [student name]?*

- ○ Detention
- ○ In-School Suspension
- ○ Out-of-School Suspension
- ○ Expulsion
- ○ None of the above

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

Survey Flow: If in-school or out-of-school suspension is selected….

How many days of suspension would you recommend?*

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

*A reminder of the school discipline policy is below:

How important were the following factors in making your disciplinary recommendation?

Survey Flow: The order of the seven factors below is randomized and broken across two pages for ease of reading

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Delivering a sufficient punishment for the student's misbehavior	○	○	○	○	○
Preventing future misbehavior from the student	○	○	○	○	○
Discouraging other students from similar behavior	○	○	○	○	○
Minimizing disruption to the learning environment	○	○	○	○	○
Ensuring the safety of other students	○	○	○	○	○
Ensuring the safety of school staff	○	○	○	○	○
Following the school discipline policy	○	○	○	○	○

End of Block: Incident 3

Start of Block: Demographics

Please indicate your gender:

- ○ Male
- ○ Female
- ○ Other

Please indicate your age:

- ○ Younger than 30
- ○ 30-39
- ○ 40-49
- ○ 50-59
- ○ 60-69
- ○ 70-79
- ○ 80 or older

Please check all races and ethnicities that apply to you:

- [ ] White or Caucasian
- [ ] Black or African American
- [ ] Hispanic or Latino
- [ ] Asian
- [ ] Middle Eastern
- [ ] Other

Please indicate your current job:

- ○ School teacher
- ○ School administrator
- ○ Other school staff (please specify)
  _____
- ○ District administrator
- ○ Other district staff (please specify)
  _____
- ○ Neither school nor district staff (please specify)
  _____

End of Block: Demographics

We thank you for your time spent taking this survey.
Your response has been recorded.

Please take a minute to share this survey with your social networks!

Simply copy the link below into your email, post, tweet, or message:
https://unc.az1.qualtrics.com/jfe/form/SV_9TDB5zy6BFzRJGJ

APPENDIX D: MTURK SURVEY PRETEST

Start of Block: Instructions

THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
COLLEGE OF ARTS AND SCIENCES
Department of Sociology

Dear MTurk Workers:

This research project is designed to collect information on the way decisions are made. You are receiving this survey because you are a survey-taker employed through Amazon Mechanical Turk. The survey should take you approximately 15 minutes to complete. The goal for the study is to have 250 people participate.

Your participation is voluntary and you may stop at any time without any penalty. Participation is anonymous; no identifying information will be collected, so participants' identities will not be known. Furthermore, individual responses will not be shared. Research results will be shared only in the aggregate (as a whole), such that no individual will be identifiable in any products of the research.

The Institutional Review Board at UNC-Chapel Hill has reviewed this study and determined it to be exempt from further consideration as it involves minimal possible risk to participants. If you have questions about the study, please contact me directly at sarahmd@live.unc.edu.

Thank you very much for your participation.

Kind regards,
Sarah M. Davis

Directions:

In the survey that follows, you will be presented with three hypothetical scenarios. You will also be provided with a response policy. With that policy in mind, please use your best judgment to recommend the best response. At the end of the survey you will be given a completion code to enter into Mechanical Turk so that you can get paid.

Please read each section carefully. You will be asked questions about what you read.

End of Block: Instructions

Start of Block: School Policy: High discretion

School Policy

Please read carefully.

In this middle school, principals and teachers have full authority as provided by law to establish and enforce standards and rules as are necessary to create orderly schools and classrooms. According to the Student Code of Conduct, disciplinary consequences may include, but are not limited to, the following:

1. Detention
2. In-school suspension
3. Out-of-school suspension
4. Expulsion

This policy will remain visible as you continue with the survey.

End of Block: School Policy: High discretion

Start of Block: School Policy: Low discretion

School Policy

Please read carefully.

In this middle school, principals and teachers have full authority as provided by law to establish and enforce standards and rules as are necessary to create orderly schools and classrooms. According to the Student Code of Conduct, disciplinary consequences are assigned to students according to their offense, as outlined below:

1. Defiance or disrespect → detention
2. Inappropriate language → detention
3. Disruption → in-school suspension for one day
4. Leaving school without permission → in-school suspension for one day
5. Skipping → in-school suspension for one day
6. Communicating threats → out-of-school suspension for three days
7. Physical aggression toward another student → out-of-school suspension for four days
8. Assault of school employee → out-of-school suspension for five days
9. Possession of drugs: first offense → out-of-school suspension for seven days
10. Possession of drugs: second offense → expulsion

This policy will remain visible as you continue with the survey.

End of Block: School Policy: Low discretion

Start of Block: Incident 1

Imagine the following...

One day in [student name]'s 4th period class, [student name] refused to do [his/her] work and disrupted other students from doing theirs. The teacher asked [him/her] to sit at the back table

Survey Flow: The order of the three following incidents is randomized for each participant.

so [he/she] could focus better, but [he/she] refused. When the teacher walked over to [student name]'s desk to talk with [him/her], [he/she] stood up and yelled, "I said I'm not going, you f***ing b****!" [He/She] then charged out of the classroom, shoving the teacher roughly out of the way, and left the building.

What disciplinary action would you recommend for [student name]?*

- ○ Detention
- ○ In-School Suspension
- ○ Out-of-School Suspension
- ○ Expulsion
- ○ None of the above

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

Survey Flow: If in-school or out-of-school suspension is selected….

How many days of suspension would you recommend?*

- 1
- 2
- 3
- 4
- 5
- 6
- 7

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

How important were the following factors in making your disciplinary recommendation?

Survey Flow: The order of the six factors below is randomized and broken across two pages for ease of reading

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Delivering a sufficient punishment for the student's misbehavior	○	○	○	○	○
Preventing future misbehavior from the student	○	○	○	○	○
Discouraging other students from similar behavior	○	○	○	○	○
Minimizing disruption to the learning environment	○	○	○	○	○
Ensuring the safety of other students	○	○	○	○	○
Ensuring the safety of school staff	○	○	○	○	○

End of Block: Incident 1

Start of Block: Incident 2

Now imagine that instead, this incident took place...

At lunch in the cafeteria, the School Resource Officer saw [student name] and another student stand up, get in each other's faces, and shout threats and insults at one another. Before the School Resource Officer could stop the fight, [student name] had pushed the other student to the ground and punched [him/her – him if the primary student is male, her if the primary student is female] in the face resulting in a bloody nose. After receiving statements from both students, it seems that the confrontation was over a shared romantic interest.

What disciplinary action would you recommend for [student name]?*

- ○ Detention
- ○ In-School Suspension
- ○ Out-of-School Suspension
- ○ Expulsion
- ○ None of the above

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

Survey Flow: If in-school or out-of-school suspension is selected….

How many days of suspension would you recommend?*

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7

---

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

How important were the following factors in making your disciplinary recommendation?

Survey Flow: The order of the six factors below is randomized and broken across two pages for ease of reading

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Delivering a sufficient punishment for the student's misbehavior	○	○	○	○	○
Preventing future misbehavior from the student	○	○	○	○	○
Discouraging other students from similar behavior	○	○	○	○	○
Minimizing disruption to the learning environment	○	○	○	○	○
Ensuring the safety of other students	○	○	○	○	○
Ensuring the safety of school staff	○	○	○	○	○

End of Block: Incident 2

Start of Block: Incident 3

Finally, imagine that instead of either of the two previous scenarios, the following occurred...

As students boarded the school bus at the end of the school day, the driver noticed [student name] had bloodshot eyes and seemed unusually spacey. The driver called the front office and when they searched [student name]'s bag they found a small bag of marijuana and a pipe. [Student name] has been caught with marijuana once already this school year.

What disciplinary action would you recommend for [student name]?*

- ○ Detention
- ○ In-School Suspension
- ○ Out-of-School Suspension
- ○ Expulsion
- ○ None of the above

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

Survey Flow: If in-school or out-of-school suspension is selected….

How many days of suspension would you recommend?*

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7

---

*A reminder of the school discipline policy is below:

Survey Flow: The participant's assigned school discipline policy is shown, either allowing for high or low discretion.

How important were the following factors in making your disciplinary recommendation?

Survey Flow: The order of the seven factors below is

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Delivering a sufficient punishment for the student's misbehavior	○	○	○	○	○
Preventing future misbehavior from the student	○	○	○	○	○
Discouraging other students from similar behavior	○	○	○	○	○
Minimizing disruption to the learning environment	○	○	○	○	○
Ensuring the safety of other students	○	○	○	○	○
Ensuring the safety of school staff	○	○	○	○	○

End of Block: Incident 3

Start of Block: Demographics

Please indicate your gender:

- ○ Male
- ○ Female
- ○ Other

Please indicate your age:

- ○ Younger than 30
- ○ 30-39
- ○ 40-49
- ○ 50-59
- ○ 60-69
- ○ 70-79
- ○ 80 or older

Please check all races and ethnicities that apply to you:

- [ ] White or Caucasian
- [ ] Black or African American
- [ ] Hispanic or Latino
- [ ] Asian
- [ ] Middle Eastern
- [ ] Other

End of Block: Demographics

Start of Block: Pretest Questions

What factors shaped your decision-making in the scenarios above?

How much freedom did you feel you had based on the school's discipline policy to choose disciplinary actions for the scenarios described?

	No freedom 0	1	2	3	4	Complete freedom 5
1						

If you had to guess, what would you say was the race/ethnicity of the student, [student name]?

- White or Caucasian
- Black or African American
- Hispanic/Latino
- Asian
- Other

If you had to guess, what would you say was the gender of the student described above?

- ○ Male
- ○ Female

If you had to guess, what would you saw was the class/socioeconomic status of the student described above?

- ○ Poor or working class
- ○ Middle income or middle class
- ○ Wealthy or upper class

What was confusing or unclear about the survey?

Any other comments you would like to add:

End of Block: Pretest Questions

APPENDIX E: MTURK NAMES PRETEST

Start of Block: Instructions

THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
COLLEGE OF ARTS AND SCIENCES
Department of Sociology

Dear MTurk Workers:

This research project is designed to collect information on the way people perceive names. You are receiving this survey because you are a survey-taker employed through Amazon Mechanical Turk. The survey should take you about 1 minute to complete. The goal for the study is to have 400 people participate.

Your participation is voluntary and you may stop at any time without any penalty. Participation is anonymous; no identifying information will be collected, so participants' identities will not be known. Furthermore, individual responses will not be shared. Research results will be shared only in the aggregate (as a whole), such that no individual will be identifiable in any products of the research.

The Institutional Review Board at UNC-Chapel Hill has reviewed this study and determined it to be exempt from further consideration as it involves minimal possible risk to participants. If you have questions about the study, please contact me directly at sarahmd@live.unc.edu.

Thank you very much for your participation.

Kind regards,
Sarah M. Davis

Directions:

In the survey that follows, you will be presented with a variety of people's names. Please use your best judgment to identify the features you associate with that name. At the end of the survey you will be given a completion code to enter into Mechanical Turk so that you can get paid.

Please read each section carefully. You will be asked questions about what you read.

End of Block: Instructions

Start of Block: Prompt

People's names are often associated with a certain race/ethnicity and a certain social class.

End of Block: Prompt

Survey Flow: Participants are given a sequence of six questions asking for either their race or class

Start of Block: Names

Survey flow: Participants are given a sequence of six questions asking for either their race or class association with six of the twelve names below.

If you had to guess, what would you say was the [race/ethnicity] of a [girl/boy] named [Hilary/Susan/Amy/Ebony/Kenya/Tyra/Cody/Dustin/Steven/Jamal/Terrell/Tremayne]?

- ○ White or Caucasian
- ○ Black or African American
- ○ Hispanic/Latino
- ○ Asian
- ○ Other

If you had to guess, what would you say was the class/socioeconomic status of a [girl/boy] named [Hilary/Susan/Amy/Ebony/Kenya/Tyra/Cody/Dustin/Steven/Jamal/Terrell/Tremayne]?

- ○ Poor or working class
- ○ Middle class
- ○ Wealthy or upper class

Start of Block: Demographics

Please indicate your gender:

- Male
- Female
- Other

Please indicate your age:

- Younger than 30
- 30-39
- 40-49
- 50-59
- 60-69
- 70-79
- 80 or older

Please check all races and ethnicities that apply to you:

- ☐ White or Caucasian
- ☐ Black or African American
- ☐ Hispanic or Latino
- ☐ Asian
- ☐ Middle Eastern
- ☐ Other

End of Block: Demographics

Start of Block: Additional Comments

Any comments you would like to share:

_____

_____

_____

_____

_____

End of Block: Additional Comments

Thank you for your time spent taking this survey.
Your response has been recorded.

Your MTurk completion code is:
${e://Field/MTurkCode}

To receive payment for participating,
click “Accept HIT” in the Mechanical Turk window,
enter this validation code, then click “Submit”.

REFERENCES

Berinsky, Adam J. Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(1):351-368.

Blake, Jamilia J., Bettie Ray Butler, Chance W. Lewis, and Alicia Darensbourg. 2010. "Unmasking the Inequitable Discipline Experiences of Urban Black Girls: Implications for Urban Education Stakeholders." *The Urban Review* 43(1):90-106.

Bobo, Lawrence. 1988. "Group Conflict, Prejudice, and the Paradox of Contemporary Racial Attitudes." Pp. 85-114 in *Eliminating Racism: Profiles in Controversy*, edited by Phyllis A. Katz and Dalmas A. Taylor. New York, NY: Springer Science+Business Media.

Bobo, Lawrence, James R. Kluegel, and Ryan A. Smith. 1997. "Laissez-Faire Racism: The Crystallization of a Kinder, Gentler Antiblack Ideology." Ch. 2 in *Racial Attitudes in the 1990s: Continuity and Change* edited by Steven A. Tuch and Kack K. Martin. Westport, CT: Praeger Publishers.

Bobo, Lawrence D. and Mia Tuan. 2006. *Prejudice in Politics: Group Position, Public Opinion, and the Wisconsin Treaty Rights Dispute*. Cambridge, MA: Harvard University Press.

Bonilla-Silva, Eduardo. 2018. *Racism Without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America.* 5th ed. New York: Rowman and Littlefield.

Collins, Patricia Hill. 2009. *Black Feminist Thought.* 2nd ed. New York, NY: Routledge.

Correll, Shelley. 2013. "Creating a Level Playing Field." Clayman Institute, Stanford University. YouTube website. Retrieved October 19, 2016 (https://www.youtube.com/watch?v=YPoymWLNjVk)

Dixon, John, Kevin Durrheim, and Manuela Thomae. 2017. "The Principle-Implementation Gap in Attitudes Towards Racial Equality (and How to Close It)." Advances in Political Psychology 38(1):91-126.

Donovan, Roxanne A. 2007. "To Blame or Not to Blame: Influences of Target Race and Observer Sex on Rape Blame Attribution." *Journal of Interpersonal Violence* 22(6):722-736.

Doyle Jr., Joseph J. 2008. "Child Protection and Adult Crime: Using Investigator Assignment to Estimate Causal Effects of Foster Care." *Journal of Political Economy* 116(4):746-770.

Eberhardt, Jennifer L. 2005. "Imagining Race." *The American Psychologist* 60(2): 181-190.

Fabelo, Tony, Michael D. Thompson, Martha Platkin, Dottie Carmichael, Miner P. Marchbanks, Eric A. Booth. 2011. *Breaking Schools' Rules: A Statewide Study of How School Discipline Relates to Students' Success and Juvenile Justice Involvement*. The Public

Policy Research Institute at Texas A&M University. New York: Council of State Governments Justice Center. Retrieved October 10, 2016 (https://csgjusticecenter.org/youth/breaking-schools-rules-report/)

Fenning, Pamela, Rose, Jennifer. 2007. "Overrepresentation of African American Students in Exclusionary Discipline: The Role of School Policy." *Urban Education* 42(6): 536-559.

Ferguson, Ann Arnette. 2000. *Bad Boys: Public Schools in the Making of Black Masculinity*. Ann Arbor: The University of Michigan Press.

Fish, Rachel Elizabeth. 2016. "The Racialized Construction of Exceptionality: Experimental Evidence of Race/Ethnicity Effects on Teachers' Interventions." *Social Science Research.* Retrieved June 30, 2017. http://dx.doi.org/10.1016/j.ssresearch.2016.08.007

Gaddis, S. Michael. 2015. "Discrimination in the Credential Society: An Audit Study of Race and College Selectivity in the Labor Market." *Social Forces* 93(4):1451-1479.

Gaddis, S. Michael. 2017. "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science* 4(1): 469-489.

Gaddis, S. Michael, and Raj Ghoshal. 2015. "Arab American Housing Discrimination, Ethnic Competition, and the Contact Hypothesis." *The ANNALS of the American Academy of Political and Social Science* 660(1): 282-299.

Gilliam, Walter S. Angela N. Maupin, Chin R. Reyes, Maria Accavitti, and Frederick Shic. 2016. "Do Early Educators' Implicit Biases Regarding Sex and Race Relate to Behavioral Expectations and Recommendations of Preschool Expulsions and Suspensions?" Retrieved October 17, 2018 (https://medicine.yale.edu/childstudy/zigler/publications/Preschool%20Implicit%20Bias%20Policy%20Brief_final_9_26_276766_5379_v1.pdf).

Haider, Adil H., Janel Sexton, N. Sriram, Lisa A. Cooper, David T. Efron, Sandra Swoboda, Cassandra V. villegas, Elliott R. Haut, Morgan Bonds, Peter J. Pronovost, Pamela A. Lipsett, Julia A. Freischlag, and Edward E. Cornwell. 2011. Association of Unconscious Race and Social Class Bias With Vignette-Based Clinical Assessments by Medical Students." Journal of the American Medical Association 306(9): 942-951.

Ioanide, Paula. 2015. *The Emotional Politics of Racism: How Feelings Trump Facts in an Era of Colorblindness*. Stanford, CA: Stanford University Press.

Kaufman, Joy S., Sarah S. Jaser, Ellen L. Vaughan, Jesse S. Reynolds, John Di Donato, Stanley N. Bernard, Maria Hernandez-Brereton. 2009. "Patterns in Office Referral Data by Grade, Race/Ethnicity, and Gender." Journal of Positive Behavior Interventions 12(1):44-54.

Levin, Irwin P. 1999. *Relating Statistics and Experimental Design.* Thousand Oaks, CA: Sage

Publications, Inc.

Lewis, Amanda E. and John B. Diamond. 2015. *Despite the Best Intentions: How Racial Inequality Thrives in Good Schools.* New York: Oxford University Press.

Losen, Daniel J., Gillespie, Jonathan. 2012. "Opportunity Suspended: The Disparate Impact of Disciplinary Exclusion from School." The Civil Rights Project/Proyecto Derechos Civiles. Retrieved October 10, 2016 (https://escholarship.org/uc/item/3g36n0c3).

Martin, Karin A. 1998. "Becoming a gendered body: Practices of preschools." *American Sociological Review* 63: 494-511.

McPherson, Miller; Lynn Smith-Lovin; and James M. Cook. 2001. "Birds Of A Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415–44.

Morris, Edward W. 2005. "'Tuck in That Shirt!' Race, Class, Gender, and Discipline in an Urban School." *Sociological Perspectives* 48(1): 25-48.

Morris, Edward W. and Brea L. Perry. 2017. "Girls Behaving Badly? Race, Gender, and Subjective Evaluation in the Discipline of African American Girls." *Sociology of Education* 90(2):127-148.

Morris, Monique W. 2016. *Pushout: The Criminalization of Black Girls in Schools*. New York: The New Press.

Omi, Michael and Howard Winant. 2014. *Racial Formation in the United States: From the 1960s to the 1990s.* 3rd ed. New York, New York: Routledge.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108(5):937-975.

Pager, Devah and Lincoln Quillian. 2005. "Walking the Talk? What Employers Say Versus What They Do." American Sociological Review 70(3):355-380.

Perry, Twila L. 1991. "Race and Child Placement: The Best Interests Test and the Cost of Discretion." *Journal of Family Law* 29(1):51-127.

Phillips, Derek L. and Kevin J. Clancy. 1970. "Response Biases in Field Studies of Mental Illness." *American Sociological Review* 35(3):503-515.

Rocque, Michael. 2010. "Office Discipline and Student Behavior: Does Race Matter?" *American Journal of Education* 116(4):557-581.

Rocque, Michael, and Raymond Paternoster. 2011. "Understanding the Antecedents of the 'School-to-Jail' Link: The Relationship Between Race and School Discipline." *The Journal of Criminal Law and Criminology* 101(2):633-665.

Rome, Dennis M., In Soo Son, and Mark S. Davis. 1995. “Police Use of Excessive Force: Does the Race of the Suspect Influence Citizens’ Perceptions?” *Social Justice Research* 8(1):41-56.

Sears, David O. 1988. “Symbolic Racism.” Pp. 53-84 in *Eliminating Racism: Profiles in Controversy*, edited by Phyllis A. Katz and Dalmas A. Taylor. New York, NY: Springer Science+Business Media.

Sears, David O., John J. Hetts, Jim Sidanius, and Lawrence Bobo. 2000. “Race in American Politics: Framing the Debates. Pp. 1-43 in *Racialized Politics: The Debate About Racism in America*. Chicago, IL: University of Chicago Press.

Singleton, Royce and Bruce C. Straits. 2018. *Approaches to Social Research*. 6th ed. New York: Oxford University Press.

Skiba, Russell J., Arredondo, Mariella I., Williams, Natasha T. 2014. “More Than a Metaphor: The Contribution of Exclusionary Discipline to a School-to-Prison Pipeline.” *Equity and Excellence in Education* 47(4):546-564.

Skiba, Russell J., Robert H. Horner, Choong-Geun Chung, M. Karega Rausch, Seth L. May, Tary Tobin. 2011. “Race is Not Neutral: A National Investigation of African American and Latino Disproportionality in School Discipline.” *School Psychology Review* 40(1):85-107.

Skiba, Russell J., Robert S. Michael, Abra Carroll Nardo, Reece L. Peterson. 2002. “The Color of Discipline: Sources of Racial and Gender Disproportionality in School Punishment.” *The Urban Review* 34(4):317-342.

Skiba, Russell, Cecil R. Reynolds, Sandra Graham, Peter Sheras, Jane Close Conoley, and Enedina Garcia-Vazquez. 2008. “Are Zero Tolerance Policies Effective in the Schools? An Evidentiary Review and Recommendations.” *American Psychologist* 63: 852–62.

Stern, Michael J., Ipek Bilgen, Colleen McClain, and Brian Hunscher. 2017. “Effective Sampling from Social Media Sites and Search Engines for Web Surveys: Demographic and Data Quality Differences in Surveys of Google and Facebook Users.” *Social Science Computer Review* 35(6): 713-732.

U.S. Department of Education Office for Civil Rights. 2014. *Civil Rights Data Collection: Data Snapshot (School Discipline)*. Issue Brief No. 1. March 21.

Vavrus, Frances and KimMarie Cole. 2002. “‘I Didn’t Do Nothin’’: The Discursive Construction of School Suspension.” *The Urban Review* 34(2):87-111.

Wallace, Jr., John M., Sara Goodkind, Cynthia M. Wallace, and Jerald G. Bachman. 2008.

"Racial, Ethnic, and Gender Differences in School Discipline among U.S. High School Students: 1991-2005." *Negro Educational Review* 59(1-2):47-62.

Weisshaar, Katherine. 2018. "From Opt Out to Blocked Out: The Challenges for Labor Market Re-Entry after Family-Related Employment Lapses." American Sociological Review 83(1): 34-60

Welch, Kelly and Allison Ann Payne. 2010. "Racial Threat and Punitive School Discipline." *Social Problems* 57(1):25-48.

Wellman, David T. 1993. *Portraits of White Racism*. 2nd ed. Cambridge, MA: Cambridge University Press.

Wodtke, Geoffrey T. 2016. "Are Smart People Less Racist? Verbal Ability, Anti-Black Prejudice, and the Principle-Policy Paradox." Social Problems 63(1):21–45.